Análise de Variância

Elsa Gonçalves

(Adaptado,Cadima,J.(2021).O Modelo Linear. ISA, ULisboa)

I.3. Análise de Variância (ANOVA) de efeitos fixos

A Regressão Linear visa modelar uma variável resposta numérica (quantitativa), à custa de uma ou mais variáveis preditoras, igualmente numéricas.

Mas uma variável resposta numérica pode depender de variáveis qualitativas (categóricas), ou seja, de um ou mais factores.

A Análise de Variância (ANOVA) é uma metodologia estatística para lidar com este tipo de situações.

A ANOVA foi desenvolvida nos anos 30 do Século XX, na Estação Experimental Agrícola de Rothamstead (Inglaterra), por R.A. Fisher.

Exemplo motivador: os lírios

Até aqui ignorou-se que os 150 lírios do conjunto de dados `iris` referem-se a 50 observações em cada uma de três diferentes espécies.



iris setosa



iris versicolor

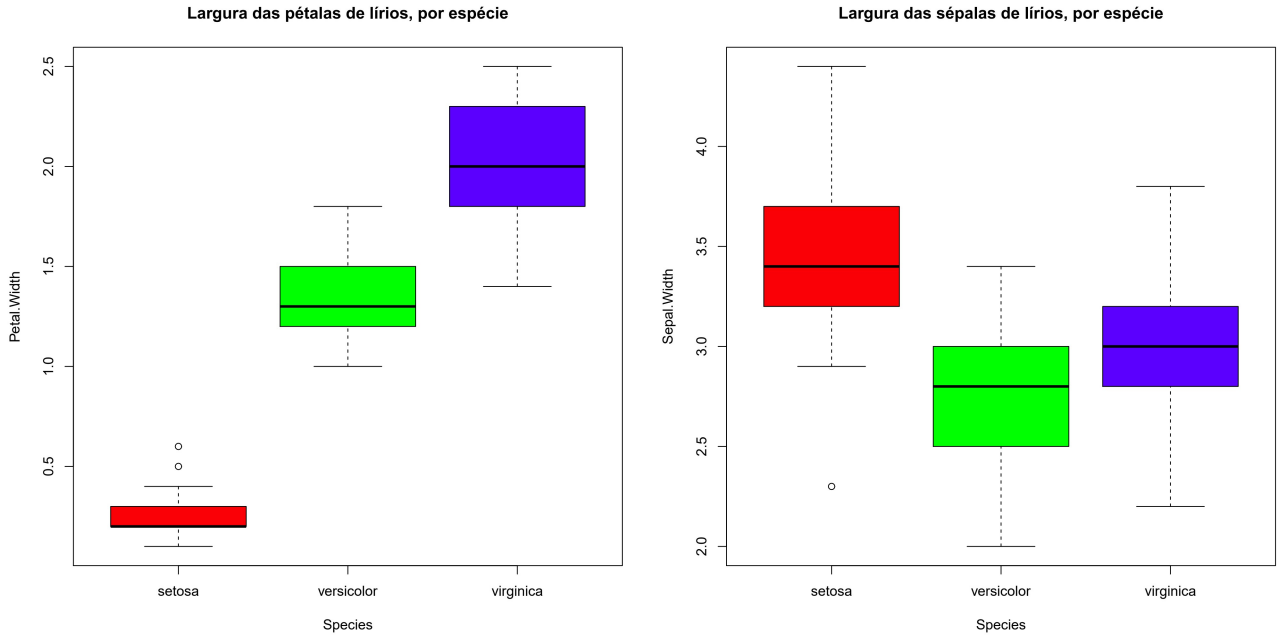


iris virginica

Poderão os valores médios de cada característica morfométrica *diferir consoante as espécies*?

Objectivo: testar a igualdade de médias duma variável, em diferentes contextos (neste exemplo, para diferentes espécies de lírios).

Dois exemplos: os lírios por espécie



As larguras das pétalas parecem diferir entre as espécies dos lírios.
As larguras das sépalas diferem menos. Eis as médias amostrais:

$$\overline{y}_{seto} = 3.428 \quad ; \quad \overline{y}_{vers} = 2.770 \quad ; \quad \overline{y}_{virg} = 2.974$$

As diferenças serão apenas um acaso da amostra?

Objectivo: Testar a igualdade das médias populacionais de cada espécie.

A ANOVA como caso particular do Modelo Linear

A Análise de Variância (ANOVA) lida com variáveis preditoras (explicativas) qualitativas. Surgiu historicamente como um método autónomo. Mas, tal como a Regressão Linear, é uma particularização do Modelo Linear.

Introduzir a ANOVA através das suas semelhanças com a Regressão Linear permite aproveitar boa parte da teoria estudada até aqui.

Terminologia

Variável resposta Y: uma variável numérica (quantitativa), que se pretende estudar e modelar.

Factor : uma variável preditora categórica (qualitativa);

Níveis do factor : as diferentes categorias ("valores") do factor, ou seja, diferentes situações experimentais onde se efectuam observações de Y.

Nos exemplos, o factor Espécie tem $k = 3$ níveis.

A ANOVA a um Factor - notação

Na ANOVA a um Factor (totalmente casualizado), a modelação da variável resposta baseia-se numa única variável preditora categórica.

Admitimos que o factor tem k níveis (no exemplo dos lírios, $k = 3$).

Admitimos que há n observações independentes de Y, sendo n_i $(i = 1, ..., k)$ correspondentes ao nível i do factor. Logo, $\sum_{i=1}^{k} n_i = n$.

Delineamentos equilibrados

No caso de igual número de observações em cada nível,

$$n_1 = n_2 = n_3 = \cdots = n_k \qquad (= n_c),$$

diz-se que estamos perante um delineamento equilibrado.

Os delineamentos equilibrados são aconselháveis (mas não obrigatórios), por várias razões que adiante se discutem.

A dupla indexação de Y

Na regressão linear indexam-se as n observações de Y com um único índice, variando de 1 a n $(\{Y_i\}_{i=1}^{n})$.

Neste novo contexto, é preferível usar dois índices para indexar as observações de Y:

- um (i) indica o nível do factor a que a observação corresponde;
- outro (j) permite distinguir as observações num mesmo nível.

Assim, a j-ésima observação de Y, no i-ésimo nível do factor, é representada por Y_{ij} , (com $i = 1, ..., k$ e $j = 1, ..., n_i$) .

A equação do modelo

A equação do modelo será mais simples do que na regressão: a única informação disponível para prever Y_{ij} é que a observação corresponde ao nível i do factor.

Não há informação no modelo para explicar diferentes valores de Y em repetições num mesmo nível do factor: será considerada variação aleatória.

Uma primeira equação do modelo é:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \qquad \text{com} \quad E[\varepsilon_{ij}] = 0 \ ,$$

onde μ_i representa o valor esperado das observações Y_{ij} efectuadas no nível i do factor: $\mu_i = E[Y_{ij}] = E[Y|\text{obs. nivel } i]$.

Uma equação para Y_{ij}

Para poder enquadrar a ANOVA na teoria do Modelo Linear já estudada, é conveniente re-escrever as médias de nível na forma:

$$E[Y_{ij}] = \mu_i = \mu + \alpha_i \ .$$

O parâmetro μ é comum a todas as observações, enquanto os parâmetros α_i são específicos para cada nível (i) do factor.
Cada α_i é designado o efeito do nível i.

Admite-se que Y_{ij} oscila aleatoriamente em torno do seu valor médio:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \ ,$$

com $E[\varepsilon_{ij}] = 0$. Mas como relacionar esta equação do modelo com um Modelo Linear?

O modelo ANOVA como um Modelo Linear

A equação geral $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, nas n_1 observações do nível $i = 1$ fica:

$$Y_{1j} = \mu + \alpha_1 + \varepsilon_{1j} \ ,$$

nas n_2 observações efectuadas no nível $i = 2$ fica:

$$Y_{2j} = \mu + \alpha_2 + \varepsilon_{2j} \ ,$$

etc.. Este conjunto de k equações pode ser escrita como uma única equação geral, que é a equação dum modelo linear:

$$Y_{ij} = \mu + \alpha_1 \mathscr{I}_{1_{ij}} + \alpha_2 \mathscr{I}_{2_{ij}} + \ldots + \alpha_k \mathscr{I}_{k_{ij}} + \varepsilon_{ij} \ ,$$

onde $\mathscr{I}_m$ é a variável indicatriz do nível m do factor:

$$\mathscr{I}_{m_{ij}} = \begin{cases} 1 & , \quad \text{se } i = m \\ 0 & , \quad \text{se } i \neq m \end{cases}$$

A relação de base em notação vectorial

Em notação matricial/vectorial, a equação de base será:

$$\begin{aligned} \vec{\mathbf{Y}} &= \mu\,\vec{\mathbf{1}}_n + \alpha_1\,\vec{\mathscr{I}}_1 + \alpha_2\,\vec{\mathscr{I}}_2 + \alpha_3\,\vec{\mathscr{I}}_3 + \ldots + \alpha_k\,\vec{\mathscr{I}}_k + \vec{\boldsymbol{\varepsilon}} \\ \Leftrightarrow \quad \vec{\mathbf{Y}} &= \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}\,, \end{aligned}$$

As colunas de **X** são: o vector $\vec{\mathbf{1}}_n$ e os vectores das indicatrizes $\vec{\mathscr{I}}_i$.
O vector dos parâmetros $\vec{\boldsymbol{\beta}}$ tem elementos: μ e os efeitos α_i.

Num exemplo com $n_1 = 3$, $n_2 = 4$ e $n_3 = 2$ observações:

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

O problema do excesso de parâmetros

Existe um problema "técnico": as colunas desta matriz **X** são linearmente dependentes (a soma das indicatrizes é o vector dos n uns) , pelo que a matriz $\mathbf{X}^t\mathbf{X}$ não é invertível. Há um excesso de parâmetros no modelo.

Soluções possíveis na equação $Y_{ij} = \mu + \alpha_1 \mathscr{I}_{1_{ij}} + \alpha_2 \mathscr{I}_{2_{ij}} + \ldots + \alpha_k \mathscr{I}_{k_{ij}} + \varepsilon_{ij}$:

1. retirar o parâmetro μ do modelo.
   - corresponde a retirar a coluna de uns da matriz **X**;
   - cada α_i equivalerá a μ_i, a média do nível;
   - não se pode generalizar a situações mais complexas;
   - mais difícil de encaixar na teoria já dada do Modelo Linear.
2. impor restrições aos parâmetros: e.g., $\sum_{i=1}^{k} \alpha_i = 0$.
   - Foi a solução clássica, ainda hoje frequente em livros de ANOVA;
   - mais difícil de encaixar na teoria geral do Modelo Linear.
3. tomar $\alpha_1 = 0$: será a solução utilizada.
   - corresponde a excluir a 1a. variável indicatriz do modelo (e de **X**);
   - permite aproveitar a teoria do Modelo Linear e é generalizável.

Cada solução tem implicações na forma de interpretar os parâmetros.

A matriz do modelo com a restrição $\alpha_1 = 0$

Com a restrição $\alpha_1 = 0$, a matriz do modelo **X** tem colunas $\vec{\mathbf{1}}_n, \vec{\mathscr{I}}_2, \ldots, \vec{\mathscr{I}}_k$.
No exemplo anterior, tem-se:

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

Agora $\mu = \mu_1$ é o valor médio das observações do nível $i = 1$:

$$\begin{array}{lcccccccl} Y_{1j} & = & \mu + \varepsilon_{1j} & \Rightarrow & \mu_1 & = & E[Y_{1j}] & = & \mu & , \forall j = 1, \ldots, n_1 \\ Y_{2j} & = & \mu + \alpha_2 + \varepsilon_{2j} & \Rightarrow & \mu_2 & = & E[Y_{2j}] & = & \mu_1 + \alpha_2 & , \forall j = 1, \ldots, n_2 \\ Y_{3j} & = & \mu + \alpha_3 + \varepsilon_{3j} & \Rightarrow & \mu_3 & = & E[Y_{3j}] & = & \mu_1 + \alpha_3 & , \forall j = 1, \ldots, n_3 \end{array}$$

Os efeitos de nível α_i

Na equação duma ANOVA a um factor (acetato 228), e com a restrição $\alpha_1 = 0$, cada α_i ($i > 1$) representa o acréscimo que transforma a média do primeiro nível na média do nível i:

$$\begin{aligned}
\alpha_1 &= 0 \\
\alpha_2 &= \mu_2 - \mu_1 \\
\alpha_3 &= \mu_3 - \mu_1 \\
&\vdots \quad \vdots \quad \vdots \\
\alpha_k &= \mu_k - \mu_1
\end{aligned}$$

A igualdade de todas as médias populacionais de nível μ_i equivale a que todos os efeitos de nível sejam nulos: $\alpha_i = 0\ , \quad \forall\, i$.

O modelo ANOVA a 1 factor para efeitos inferenciais

Para completar o modelo ANOVA a um factor, admite-se que os erros aleatórios ε_{ij} têm as mesmas propriedades que numa regressão linear:

Modelo ANOVA a um factor, com k níveis

Existem n observações, Y_{ij}, das quais n_i correspondem ao nível i $(i = 1, \ldots, k)$ do factor. Tem-se:

1. $Y_{ij} = \mu_1 + \alpha_i + \varepsilon_{ij}$, $\forall i=1,\ldots,k$, $\forall j=1,\ldots,n_i$ $(\alpha_1 = 0)$.
2. $\varepsilon_{ij} \frown \mathcal{N}(0, \sigma^2)$, $\forall i,j$
3. $\{\varepsilon_{ij}\}_{i,j}$ v.a.s independentes.

O modelo tem k parâmetros: a média de Y no primeiro nível do factor, μ_1, e os acréscimos α_i $(i > 1)$ que geram as médias de cada um dos $k-1$ restantes níveis do factor. Ou seja,

$$\vec{\boldsymbol{\beta}} = (\mu_1, \alpha_2, \alpha_3, \cdots, \alpha_k)^t .$$

O modelo ANOVA a um factor - notação vectorial

De forma equivalente, em notação vectorial,

Modelo ANOVA a um factor - notação vectorial

O vector $\vec{\mathbf{Y}}$ das n observações verifica:

1. $\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{1}}_n + \alpha_2 \vec{\mathscr{I}}_2 + \alpha_3 \vec{\mathscr{I}}_3 + \ldots + \alpha_k \vec{\mathscr{I}}_k + \vec{\boldsymbol{\varepsilon}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}$, sendo
   - $\vec{\mathbf{1}}_n$ o vector de n uns e $\vec{\mathscr{I}}_2$, $\vec{\mathscr{I}}_3$, ..., $\vec{\mathscr{I}}_k$ as variáveis indicatrizes dos níveis indicados;
   - $\mathbf{X} = \left[\vec{\mathbf{1}}_n \mid \vec{\mathscr{I}}_2 \mid \vec{\mathscr{I}}_3 \mid \cdots \mid \vec{\mathscr{I}}_k \right]$ a matriz $n \times k$ do modelo; e
   - $\vec{\boldsymbol{\beta}} = (\mu_1, \alpha_2, \alpha_3, \cdots, \alpha_k)^t$ o vector dos parâmetros.
2. $\vec{\boldsymbol{\varepsilon}} \frown \mathscr{N}_n(\vec{\mathbf{0}}, \sigma^2 \mathbf{I}_n)$, sendo $\mathbf{I}_n$ a matriz identidade $n \times n$.

Trata-se de um modelo análogo a um modelo de Regressão Linear Múltipla, diferindo apenas na natureza das variáveis preditoras, que são aqui variáveis indicatrizes dos níveis 2 a k do factor.

O teste aos efeitos do factor

A hipótese de que nenhum dos níveis do factor afecte a média da variável resposta corresponde à hipótese

$$\begin{aligned} & \alpha_2 = \alpha_3 = \ldots = \alpha_k = 0 \\ \Leftrightarrow \quad & \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k \end{aligned}$$

Dado o paralelismo com os modelos de Regressão Linear, esta hipótese corresponde a dizer que todos os coeficientes das "variáveis preditoras" (na ANOVA, as variáveis indicatrizes $\vec{\mathscr{I}}_i$) são nulos.

É possível testar esta hipótese, através dum teste F de ajustamento global do modelo (ver acetato **??**) que, no contexto, chamamos Teste F aos efeitos do factor.

O Teste F aos efeitos do factor numa ANOVA

Muda-se a designação de QMR para QMF (Quadrado Médio do Factor):

Teste F aos efeitos do factor

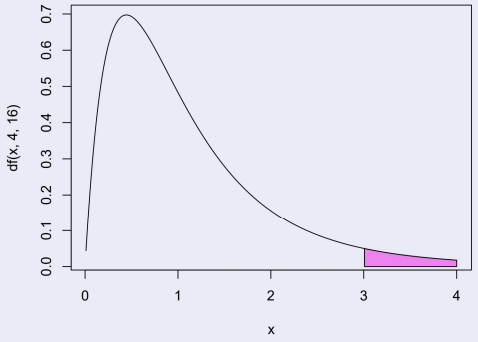
Hipóteses: $H_0 : \alpha_i = 0 \quad \forall i=2,\ldots,k$ vs. $H_1 : \exists i=2,..,k \text{ t.q. } \alpha_i \neq 0$.
[FACTOR NÃO AFECTA] vs. [FACTOR AFECTA Y]

Estatística do Teste: $F = \frac{QMF}{QMRE} \frown F_{(k-1,n-k)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rej. H_0 se $F_{calc} > f_{\alpha(k-1,n-k)}$



Notação e graus de liberdade

Neste contexto, existem fórmulas simples para algumas quantidades.

Numa ANOVA a um factor, usamos SQF, em vez de SQR, para indicar a Soma de Quadrados associada aos efeitos do Factor, embora a sua definição seja idêntica (numerador da variância dos valores ajustados).

Numa ANOVA a um factor, o número de preditores do modelo (as variáveis indicatrizes dos níveis $2, 3, \ldots, k$) é $p = k - 1$ e o número de parâmetros do modelo é $p + 1 = k$. Logo, os graus de liberdade associados a cada Soma de Quadrados são:

SQxx	g.l.
SQF	$k - 1$
SQRE	$n - k$

Os Quadrados Médios continuam a ser os quocientes das Somas de Quadrados a dividir pelos respectivos graus de liberdade.

Estimadores de parâmetros na ANOVA a um factor

Na ANOVA a um factor, as k colunas de $\mathbf{X}$ são os vectores $\vec{\mathbf{1}}_n$, $\vec{\mathscr{I}}_2$, $\vec{\mathscr{I}}_3$, ... , $\vec{\mathscr{I}}_k$. A matriz identifica as observações de cada nível do factor.

Dada a natureza especial da matriz $\mathbf{X}$, a fórmula dos parâmetros ajustados, $\hat{\vec{\boldsymbol{\beta}}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}$ gera estimadores dos parâmetros populacionais que são as quantidades amostrais análogas. Sendo $\overline{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ a média amostral das n_i observações de Y no nível i, tem-se:

$$
\begin{array}{rcll}
 & \mu_1 & \longrightarrow & \hat{\mu}_1 = \overline{Y}_{1\cdot} \\
\alpha_2 = & \mu_2 - \mu_1 & \longrightarrow & \hat{\alpha}_2 = \overline{Y}_{2\cdot} - \overline{Y}_{1\cdot} \\
\alpha_3 = & \mu_3 - \mu_1 & \longrightarrow & \hat{\alpha}_3 = \overline{Y}_{3\cdot} - \overline{Y}_{1\cdot} \\
 & \vdots & & \vdots \quad \vdots \quad \vdots \\
\alpha_k = & \mu_k - \mu_1 & \longrightarrow & \hat{\alpha}_k = \overline{Y}_{k\cdot} - \overline{Y}_{1\cdot}
\end{array}
$$

Os valores ajustados $\hat{Y}_{ij}$

Valores ajustados $\hat{Y}_{ij}$

Do que foi visto, decorre que qualquer observação tem valor ajustado igual à média amostral das observações do seu nível:

$$\hat{Y}_{ij} \quad = \quad \underbrace{\hat{\mu}_1 + \hat{\alpha}_i}_{=\hat{\mu}_i} \quad = \quad \overline{Y}_{1.} + (\overline{Y}_{i.} - \overline{Y}_{1.}) \quad = \quad \overline{Y}_{i.} \; .$$

Os valores ajustados $\hat{Y}_{ij}$ são iguais para todas as observações num mesmo nível i do factor. Tal como na Regressão, estes valores resultam de projectar ortogonalmente o vector $\vec{\mathbf{Y}}$ dos valores observados da variável resposta, sobre o subespaço $\mathscr{C}(\mathbf{X}) \subset \mathbb{R}^n$ gerado pelas colunas da matriz $\mathbf{X}$: $\vec{\hat{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$.

Numa ANOVA a um factor, o subespaço $\mathscr{C}(\mathbf{X})$ tem natureza especial: todos os vectores de $\mathscr{C}(\mathbf{X})$ têm de ter valor igual nas posições correspondentes a observações dum mesmo nível do factor.

Os resíduos e *SQRE*

Vimos que $\hat{Y}_{ij} = \hat{\mu}_i = \overline{Y}_{i\cdot}$.

O resíduo da observação Y_{ij} é dado pela sua diferença em relação à média amostral de nível:

$$E_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \overline{Y}_{i\cdot} ,$$

A Soma de Quadrados dos Resíduos é dada por:

$$SQRE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} E_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2 = \sum_{i=1}^{k} (n_i - 1) S_i^2 ,$$

onde $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2$ é a variância amostral das n_i observações de Y no i-ésimo nível do factor.

SQRE mede variabilidade no seio dos k níveis.

Fórmulas para delineamentos equilibrados

No caso de um delineamento equilibrado, i.e., $n_1 = n_2 = \ldots = n_k (= n_c)$ tem-se $n = n_c \cdot k$, e:

$$SQRE = (n_c - 1) \sum_{i=1}^{k} S_i^2$$

$$QMRE = \frac{n_c - 1}{n - k} \sum_{i=1}^{k} S_i^2 = \frac{n_c - 1}{k\,(n_c - 1)} \sum_{i=1}^{k} S_i^2 = \frac{1}{k} \sum_{i=1}^{k} S_i^2 \,.$$

Assim, em delineamentos equilibrados, o Quadrado Médio Residual é a média (simples) das k variâncias de nível da variável resposta Y.

Em delineamentos não equilibrados, o QMRE é uma média ponderada dos S_i^2 (tendo cada parcela o peso $n_i - 1$).

A Soma de Quadrados associada ao Factor

A Soma de Quadrados associada à Regressão toma, neste contexto, a designação Soma de Quadrados associada ao Factor e será representada por *SQF*. Sendo $\overline{Y}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$ a média da totalidade das n observações, tem-se:

$$SQF = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\hat{Y}_{ij} - \overline{Y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2$$

$$\Leftrightarrow \quad SQF = \sum_{i=1}^{k} n_i \left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2$$

SQF mede variabilidade entre as médias amostrais de cada nível.

Fórmulas para delineamentos equilibrados

No caso de um delineamento equilibrado $n_1 = n_2 = ... = n_k (= n_c)$,

$$SQF \quad = \quad n_c \sum_{i=1}^{k} (\overline{Y}_{i.} - \overline{Y}_{..})^2 \quad = \quad n_c (k-1) \cdot S^2_{\overline{Y}_{i..}} \quad ,$$

onde $S^2_{\overline{Y}_{i..}} = \frac{1}{k-1} \sum_{i=1}^{k} (\overline{Y}_{i.} - \overline{Y}_{..})^2$ indica a variância amostral das k médias de nível amostrais.

$$QMF \quad = \quad \frac{SQF}{k-1} \quad = \quad n_c \cdot S^2_{\overline{Y}_{i..}} \quad .$$

Assim, em delineamentos equilibrados, o Quadrado Médio associado aos efeitos do Factor, *QMF*, é proporcional à variância das k médias de nível da variável Y.

A relação entre Somas de Quadrados

A relação fundamental entre as três Somas de Quadrados (mesmo com delineamentos não equilibrados) tem um significado particular:

$$\begin{array}{ccccc} SQT & = & SQF & + & SQRE \\ \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{..})^2 & = & \sum_{i=1}^{k} n_i(\overline{Y}_{i.}-\overline{Y}_{..})^2 & + & \sum_{i=1}^{k}(n_i-1)S_i^2 \ . \end{array}$$

onde:

- $SQT = (n-1)s_y^2$ mede a variabilidade total das n observações de Y;
- SQF mede a variabilidade entre diferentes níveis do factor (variabilidade inter-níveis);
- $SQRE$ mede a variabilidade no seio dos níveis - e que portanto não é explicada pelo factor (variabilidade intra-níveis).

Esta é a origem histórica do nome "Análise da Variância": a variância de Y é decomposta ("analisada") em parcelas, associadas a diferentes causas. Aqui, as causas podem ser o efeito do factor ou outras não explicadas pelo modelo (residuais).

O quadro de síntese da ANOVA a 1 Factor

Pode-se coleccionar esta informação numa tabela-resumo da ANOVA:

Fonte	g.l.	SQ	QM	f_{calc}
Factor	$k-1$	$SQF = \sum_{i=1}^{k} n_i \cdot (\overline{y}_{i.} - \overline{y}_{..})^2$	$QMF = \frac{SQF}{k-1}$	$\frac{QMF}{QMRE}$
Resíduos	$n-k$	$SQRE = \sum_{i=1}^{k} (n_i - 1)\, s_i^2$	$QMRE = \frac{SQRE}{n-k}$	
Total	$n-1$	$SQT = (n-1)\, s_y^2$	–	–

Factores no R

O R tem uma estrutura de dados específica para variáveis qualitativas (categóricas), designada factor , criado pelo comando `factor`, aplicado a um vector contendo os nomes dos vários níveis:

```
> factor(c("Adubo 1", "Adubo 1", ...  , "Adubo 5"))
```

NOTA: Explore o comando `rep` para criar repetições de valores.

Factores no R

No objecto `iris`, a coluna `Species` é um factor. A função `summary`, com factores, devolve o número de observações em cada nível

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

ANOVAs a um Factor no R

Para efectuar uma ANOVA a um Factor no R, convém organizar os dados numa `data.frame` com duas colunas:

1. uma para os valores (numéricos) da variável resposta;
2. outra para o factor (com a indicação dos seus níveis).

As fórmulas usadas no `R` para especificar uma ANOVA a um factor são semelhantes às da regressão linear, indicando o factor como variável preditora. O `R` cria as variáveis indicatrizes necessárias.

Fórmulas para ANOVAs no `R`

Para efectuar uma ANOVA de larguras das pétalas sobre espécies, nos dados dos $n = 150$ lírios, a fórmula é:

$$\texttt{Petal.Width} \quad \sim \quad \texttt{Species}$$

uma vez que a *data frame* `iris` contém uma coluna de nome `Species` que foi definida como factor.

ANOVAs a um factor no R (cont.)

Embora seja possível usar o comando `lm` para efectuar uma ANOVA (a ANOVA é caso particular do Modelo Linear), o comando `aov` organiza a informação da forma mais tradicional numa ANOVA.

Uma ANOVA com os lírios

Eis a ANOVA da largura de pétalas sobre espécies, nos lírios:

```
> aov(Petal.Width ~ Species, data=iris)

Call:
   aov(formula = Petal.Width ~ Species, data = iris)
Terms:
                  Species Residuals
Sum of Squares  80.41333   6.15660
Deg. of Freedom         2       147

Residual standard error: 0.20465
```

ANOVAs a um factor no R (cont.)

A função `summary` também pode ser aplicada ao resultado de uma ANOVA, produzindo o quadro-resumo completo da ANOVA.

ANOVA da largura das sépalas

Eis o resultado da ANOVA do segundo exemplo do acetato 223:

```
> iris.aov <- aov(Sepal.Width ~ Species , data=iris)
> summary(iris.aov)
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  11.35   5.672   49.16 <2e-16 ***
Residuals   147  16.96   0.115
```

Neste caso, rejeita-se claramente a hipótese de que os acréscimos de nível, α_i, sejam todos nulos, pelo que se rejeita a hipótese de larguras médias de sépalas iguais em todas as espécies. Conclusão: o factor (espécie) afecta a variável resposta (largura da sépala).

A exploração ulterior de H_1

A Hipótese Nula, no teste F numa ANOVA a 1 Factor, afirma que todos os níveis do factor têm efeito nulo, isto é, que a média da variável resposta Y é igual nos k níveis do Factor:

$$\begin{aligned} & \alpha_2 = \alpha_3 = \ldots = \alpha_k = 0 \\ \Leftrightarrow \quad \mu_1 = & \mu_2 = \mu_3 = \cdots = \mu_k \end{aligned}$$

A Hipótese Alternativa diz que pelo menos um dos níveis do factor tem uma média de Y diferente do primeiro nível:

$$\begin{aligned} & \exists i \quad \text{tal que} \quad \alpha_i \neq 0 \\ \Leftrightarrow \quad & \exists i \quad \text{tal que} \quad \mu_1 \neq \mu_i \end{aligned}$$

Ou seja, nem todas as médias de nível de Y são iguais

A exploração ulterior de H_1 (cont.)

Caso se opte pela Hipótese Alternativa, fica em aberto (excepto quando $k = 2$) a questão de saber quais os níveis do factor cujas médias diferem entre si.

Mesmo com $k = 3$, a rejeição de H_0 pode dever-se a:

$$\begin{aligned}
\mu_1 = \mu_2 \neq \mu_3 \quad & \text{i.e.,} \quad \alpha_2 = 0\ ;\ \alpha_3 \neq 0 \\
\mu_1 = \mu_3 \neq \mu_2 \quad & \text{i.e.,} \quad \alpha_3 = 0\ ;\ \alpha_2 \neq 0 \\
\mu_1 \neq \mu_2 = \mu_3 \quad & \text{i.e.,} \quad \alpha_2 = \alpha_3 \neq 0; \\
\mu_i \text{ todos diferentes} \quad & \text{i.e.,} \quad \alpha_2 \neq \alpha_3 \text{ e } \alpha_2, \alpha_3 \neq 0.
\end{aligned}$$

Como optar entre estas diferentes alternativas?

A exploração ulterior de H_1 (cont.)

Podem efectuar-se testes *t-Student* aos α_is, com base na teoria já estudada anteriormente (recorde-se que um modelo ANOVA é um modelo linear).

Mas quanto maior for k, mais sub-hipóteses alternativas existem, mais testes haverá para fazer.

A multiplicação do número de testes faz perder o controlo do nivel de significância α global para o conjunto de todos os testes.

Testes de hipóteses alternativos, relativos a todas as diferenças $\mu_i - \mu_j$ de pares de médias populacionais de Y, permitem controlar o nível de significância global α do conjunto dos testes. Tais testes chamam-se testes de comparações múltiplas de médias.

As comparações múltiplas

O nível de significância α nos testes de comparação múltipla é a probabilidade de rejeitar qualquer das hipóteses $\mu_i = \mu_j$, caso todas sejam verdade, ou seja, é um nível de significância global.

Alternativamente, podem-se construir intervalos de confiança para cada diferença $\mu_i - \mu_j$, com um nível $(1 - \alpha) \times 100\%$ de confiança de que os verdadeiros valores de $\mu_i - \mu_j$ pertencem a todos os intervalos.

A mais frequente abordagem de comparações múltiplas leva o nome de Tukey, embora em rigor só seja válido para delineamentos equilibrados.

Testes de Tukey na ANOVA a um factor

Dado um delineamento a um factor, equilibrado.

Teste de Tukey às diferenças de médias de nível

Hipóteses: $H_0 : \mu_i = \mu_j, \forall\, i,j$ vs. $H_1 : \exists i,j \text{ t.q. } \mu_i \neq \mu_j$.
[FACTOR NÃO AFECTA] vs. [FACTOR AFECTA Y]

Nível de significância (global) do teste: α

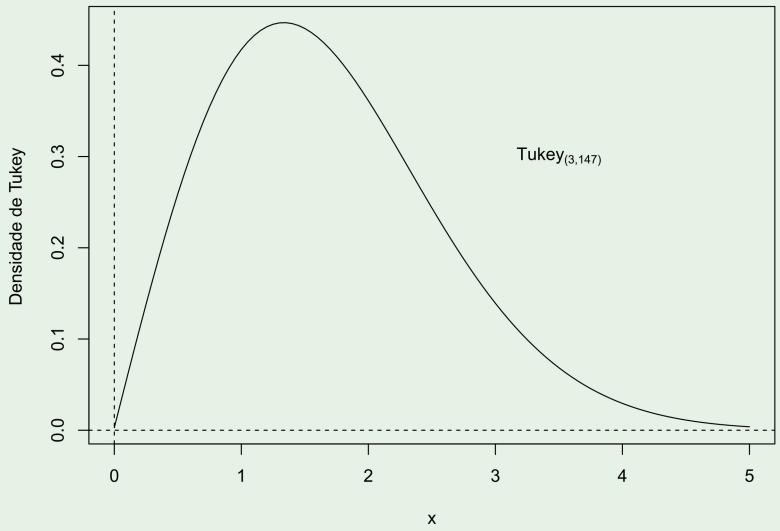
Regra: Rejeitar $\mu_i = \mu_j$ se $\left|\overline{Y}_{i.} - \overline{Y}_{j.}\right| > q_{\alpha(k,n-k)} \sqrt{\frac{QMRE}{nc}}$,
sendo $q_{\alpha(k,n-k)}$ o valor que numa distribuição de Tukey com parâmetros k e $n-k$, deixa à direita uma região de probabilidade α.

O teste permite não apenas rejeitar H_0 globalmente, como identificar o(s) par(es) de níveis (i,j) responsáveis pela rejeição (a diferença das respectivas médias amostrais excede o termo de comparação), permitindo assim conclusões sobre diferenças significativas em cada par de médias.

Distribuição de Tukey

Distribuição Tukey na ANOVA a um factor: lírios

Eis a função densidade da distribuição de Tukey, correspondente ao exemplo dos lírios, com $k = 3$ e $n - k = 147$:



Na *webpage* da disciplina encontra-se uma tabela da distribuição de Tukey.

Intervalos de Confiança para $\mu_i - \mu_j$

Alternativamente, podem construir-se intervalos de confiança para todas as diferenças de pares de médias de nível, $\mu_i - \mu_j$, com um grau de confiança global $(1-\alpha)\times 100\%$.

Concretamente, tem-se $(1-\alpha)\times 100\%$ de confiança em como todas as diferenças de médias de nível $\mu_i - \mu_j$ estão em intervalos da forma:

$$\left] \quad (\overline{y}_{i\cdot} - \overline{y}_{j\cdot}) - q_{\alpha(k,n-k)}\sqrt{\frac{QMRE}{n_c}} \quad , \quad (\overline{y}_{i\cdot} - \overline{y}_{j\cdot}) + q_{\alpha(k,n-k)}\sqrt{\frac{QMRE}{n_c}} \quad \right[$$

Se para qualquer par (i,j) de níveis, o intervalo correspondente não contém o valor zero, então $\mu_i = \mu_j$ não é admissível.

Comparações Múltiplas de Médias no R

As comparações múltiplas de médias de nível, com base no resultado de Tukey, podem ser facilmente efectuadas no R.

O termo de comparação nos testes a $\mu_i - \mu_j = 0$ é $q_{\alpha(k,n-k)} \cdot \sqrt{\frac{QMRE}{n_c}}$.

Os quantis $q_{\alpha(k,n-k)}$ duma distribuição de Tukey são calculados no R, através da função `qtukey`.

O quantil de ordem $1-\alpha$ na distribuição de Tukey obtém-se assim:

```
> qtukey(1-α, k, n-k)
```

O valor de $\sqrt{QMRE}$ é dado pelo comando `aov`, sob a designação "*Residual standard error*".

Comparações Múltiplas de Médias no

O comando `TukeyHSD` calcula os intervalos de confiança a $(1-\alpha)\times 100\%$ para as diferenças de médias.

Tukey nos lírios

```
> TukeyHSD(aov(Sepal.Width ~ Species, data=iris))
  Tukey multiple comparisons of means
    95% family-wise confidence level
$Species
                       diff         lwr         upr     p adj
versicolor-setosa    -0.658 -0.81885528 -0.4971447 0.0000000
virginica-setosa     -0.454 -0.61485528 -0.2931447 0.0000000
virginica-versicolor  0.204  0.04314472  0.3648553 0.0087802
```

O intervalo a 95% de confiança para $\mu_2 - \mu_1$ (`versicolor-setosa`) é

$$]\, -0.8189\, ,\, -0.4971\, [\, .$$

Nenhum dos intervalos inclui o valor zero, concluindo-se que $\mu_i \neq \mu_j$, para qualquer $i \neq j$, ou seja, todas as médias de espécie são diferentes.

Comparações Múltiplas de Médias no R (cont.)

O valor de prova indicado (`p adj`) é o menor valor de α para o qual uma dada diferença de médias, $\overline{y}_{i.} - \overline{y}_{j.}$, seria considerada não significativa.

Tukey nos lírios (cont.)

```
> TukeyHSD(aov(Sepal.Width ~ Species, data=iris))
  Tukey multiple comparisons of means
    95% family-wise confidence level
$Species
                       diff          lwr        upr     p adj
versicolor-setosa    -0.658 -0.81885528 -0.4971447 0.0000000
virginica-setosa     -0.454 -0.61485528 -0.2931447 0.0000000
virginica-versicolor  0.204  0.04314472  0.3648553 0.0087802
```

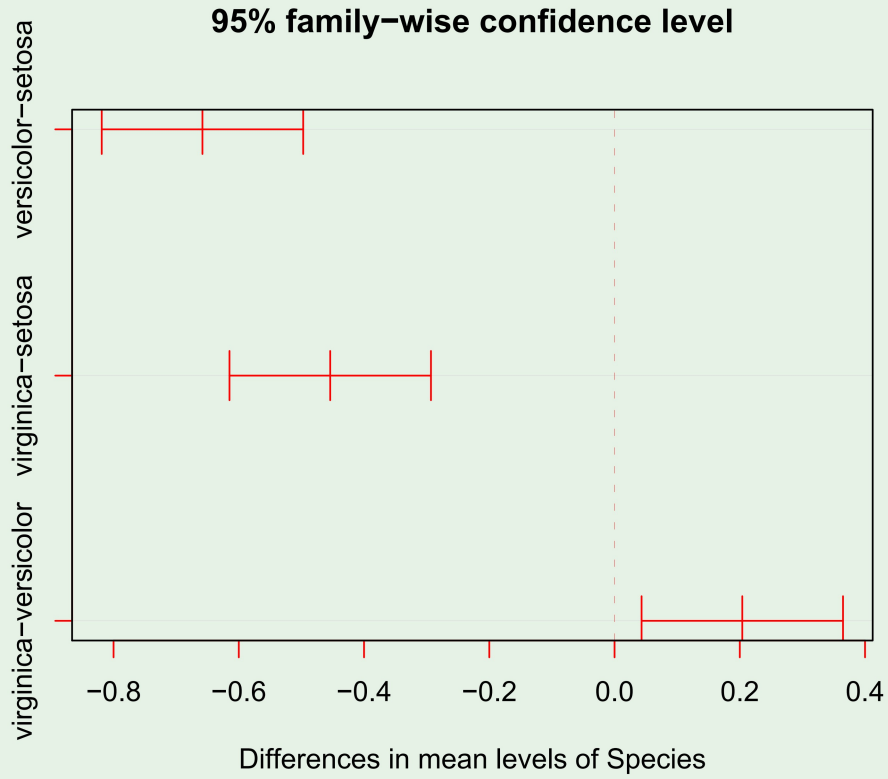
Assim, para $\alpha \leq 0.00878$, a diferença de médias amostrais para as espécies *virginica* e *versicolor* já seria considerada não significativa. Ou seja, apenas intervalos com mais de $(1-\alpha) \times 100\% = 99.122\%$ de confiança para essa diferença de médias conteriam o valor zero.

Representação gráfica das comparações múltiplas

A função `plot`, aplicada ao resultado da função `TukeyHSD`, permite visualizar os intervalos de confiança para as comparações das médias de nível.

Tukey nos lírios (cont.)

```
> plot(TukeyHSD(aov(Sepal.Width ~ Species, data=iris)))
```



Representação gráfica das comparações múltiplas

Usando library(agricolae) e a função `HSD.test`, também se obtêm as comparações das médias de nível.

Tukey nos lírios (cont.)

```
> iris.aov<-aov(Sepal.Width ~ Species, data=iris)
> library (agricolae)
> HSD.test(iris.aov, "Species",console=TRUE)

--
Critical Value of Studentized Range: 3.348424

Minimun Significant Difference: 0.1608553

Treatments with the same letter are not significantly different.

           Sepal.Width groups
setosa           3.428      a
virginica        2.974      b
versicolor       2.770      c
```

Delineamentos não equilibrados

Quando o delineamento da ANOVA a um Factor não é equilibrado (isto é, existe diferente número de observações nos vários níveis do factor), os teste/ICs de Tukey agora enunciados não são, em rigor, válidos.

Mas, para delineamentos em que o desequilíbrio no número de observações não seja muito acentuado, é possível um resultado aproximado, que a função `TukeyHSD` do  incorpora.

Análise de Resíduos na ANOVA a 1 Factor

A validade dos pressupostos do modelo estuda-se de forma idêntica ao que foi visto na Regressão Linear, tal como os diagnósticos para observações especiais. Mas há algumas particularidades.

Numa ANOVA a um factor, os resíduos aparecem empilhados em k colunas nos gráficos de e_{ij} vs. $\hat{y}_{ij}$, porque qualquer valor ajustado $\hat{y}_{ij} = \overline{y}_{i.}$ é igual para observações num mesmo nível do factor.

Este padrão não corresponde a qualquer violação dos pressupostos do modelo.

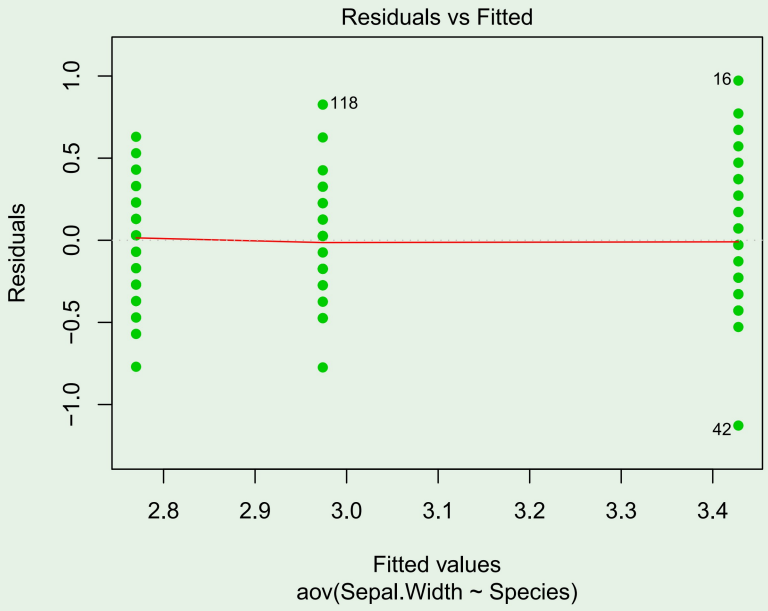
Por outro lado, todas as observações dum mesmo nível do factor terão idêntico efeito alavanca, igual a $\frac{1}{n_i}$. Sobretudo no caso de delineamentos equilibrados, isto torna os gráficos de efeitos alavanca pouco úteis neste contexto.

Análise de Resíduos na ANOVA a 1 Factor (cont.)

Padrão de resíduos numa ANOVA a 1 Factor.

Gráfico de resíduos nos lírios

```
> plot(aov(Sepal.Width ~ Species, data=iris), which=1, pch=16)
```



Estes gráficos continuam a ser úteis para validar o pressuposto de homogeneidade de variâncias dos erros aleatórios.

Violações aos pressupostos da ANOVA

As n_i repetições em cada um dos k níveis do factor, permitem testar formalmente se as variâncias dos erros aleatórios diferem entre os níveis do factor (testes de Bartlett ou de Levene, que não são dados).

Violações aos pressupostos do modelo não têm sempre igual gravidade. Alguns comentários gerais:

- O teste F da ANOVA e as comparações múltiplas de Tukey são relativamente robustos a desvios à hipótese de normalidade.
- As violações ao pressuposto de variâncias homogéneas são em geral menos graves no caso de delineamentos equilibrados, mas podem ser graves em delineamentos não equilibrados.
- A falta de independência entre erros aleatórios é a violação mais grave dos pressupostos e deve ser evitada, o que é em geral possível com um delineamento experimental adequado.

Uma advertência

Na formulação clássica do modelo ANOVA a um Factor, e a partir da equação-base

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} , \qquad \forall\, i,j$$

em vez de impor a condição $\alpha_1 = 0$, impõe-se a condição $\sum_i \alpha_i = 0$.

Esta condição alternativa:

- Muda a forma de interpretar os parâmetros (μ é agora uma espécie de média geral de Y e α_i o desvio da média do nível i em relação a essa média geral);
- Muda os estimadores dos parâmetros.
- Não muda o resultado do teste F à existência de efeitos do factor, nem a qualidade global do ajustamento.

Delineamentos factoriais a dois factores

Vamos agora considerar delineamentos experimentais com dois factores.

A existência de mais do que um factor pode resultar de:

- pretender-se realmente estudar eventuais efeitos de mais do que um factor sobre a variável resposta;
- a tentativa de controlar a variabilidade experimental.

Historicamente, à segunda situação corresponde a designação blocos. Na primeira fala-se apenas em factores. Mas são situações análogas.

Um exemplo

Pretende-se analisar o rendimento de 5 diferentes variedades de trigo.
Os rendimentos são também afectados pelos tipo de solos usados.

Nem sempre é possível ter terrenos homogéneos numa experiência.
Mesmo que seja possível, pode não ser desejável, por se limitar a validade dos resultados a um único tipo de solos.

Admita-se que estamos interessados em quatro terrenos, com solos diferentes. Cada terreno pode ser dividido em cinco parcelas viáveis para o trigo, tendo-se ao todo 20 parcelas.

Em vez de repartir aleatoriamente as 5 variedades pelas 20 parcelas, é preferível forçar cada tipo de terreno a conter uma parcela com cada variedade. Apenas dentro dos terrenos haverá casualização.

Um exemplo (cont.)

A situação descrita no acetato anterior é a seguinte:

Terreno 1	Var.1	Var.3	Var.4	Var.5	Var.2
Terreno 2	Var.4	Var.3	Var.5	Var.1	Var.2
Terreno 3	Var.2	Var.4	Var.1	Var.3	Var.5
Terreno 4	Var.5	Var.2	Var.4	Var.1	Var.3

Houve uma restrição à casualização total: dentro de cada terreno há casualização, mas obriga-se cada terreno a ter uma parcela associada a cada nível do factor variedade.

A situação agora descrita corresponde a ter introduzido um segundo factor, o factor terreno. Neste exemplo temos um delineamento factorial a dois factores (*two-way ANOVA*), sendo um dos factores a variedade de trigo e o outro o tipo de solos.

Representação delineamento factorial (2 factores)

Um delineamento factorial é um delineamento em que há observações para todas as possíveis combinações de níveis de cada factor.

		Factor B				
	Níveis	B_1	B_2	B_3	...	B_b
FACTOR A	A_1	× × ×	× × ×	× × ×	...	× × ×
	A_2	× × ×	× × ×	× × ×	...	× × ×
	A_3	× × ×	× × ×	× × ×	...	× × ×
	⋮	⋮	⋮	⋮	⋱	⋮
	A_a	× × ×	× × ×	× × ×	...	× × ×

Atenção: Esta esquematização não corresponde a qualquer organização espacial.

Célula: cruzamento dum nível dum Factor com um nível do outro Factor. Corresponde a uma situação experimental. Nesta esquematização, há *ab* células, cada uma com 3 observações.

Modelos ANOVA a 2 Factores: notação

Admita-se a existência de:

- Uma variável resposta Y;
- Um Factor A, com a níveis;
- Um Factor B, com b níveis;
- n observações, com pelo menos uma em cada uma das ab situações experimentais (células).

O número de observações na célula correspondente ao nível i do factor A, e j do factor B é representado por n_{ij}.

O número total de observações é: $n = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij}$.

Notação

Cada observação da variável resposta é identificada com três índices,

$$Y_{ijk}$$

onde:

- i indica o nível i do Factor A $\quad (i = 1, 2, ..., a)$.
- j indica o nível j do Factor B $\quad (j = 1, 2, ..., b)$.
- k indica a repetição k na célula (i, j) $\quad (k = 1, 2, ..., n_{ij})$.

Delineamento equilibrado

Se o número de observações for igual em todas as células, $n_{ij} = n_c$, $\forall i, j$, estamos perante um delineamento equilibrado.

Estudaremos dois diferentes modelos ANOVA para um delineamento factorial com 2 factores.

Modelo ANOVA a 2 factores (sem interacção)

Um primeiro modelo prevê a existência de dois diferentes tipos de efeitos associados aos níveis de cada factor. Admite-se que o valor esperado de cada observação Y_{ijk} é da forma:

$$E[Y_{ijk}] \quad = \quad \mu_{ij} \quad = \quad \mu + \alpha_i + \beta_j \quad , \quad \forall i,j,k.$$

O parâmetro μ é comum a todas as observações.

Cada parâmetro α_i é um acréscimo que pode diferir entre níveis do Factor A, e é designado o efeito do nível i do factor A.

Cada parâmetro β_j é um acréscimo que pode diferir entre níveis do Factor B, e é designado o efeito do nível j do factor B.

Admite-se que todos estes parâmetros são constantes.

Admite-se que a variação de Y_{ijk} em torno do seu valor médio é aleatória e dada por um erro aleatório aditivo, ε_{ijk} (com $E[\varepsilon_{ijk}] = 0$):

$$Y_{ijk} \; = \; \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \; ,$$

As variáveis indicatrizes de nível de cada factor

A equação de base do modelo ANOVA a 2 factores (sem interacção) também pode ser escrita na forma vectorial, recorrendo a variáveis indicatrizes de pertença a cada nível de cada factor.

$\vec{\mathbf{Y}}$ o vector aleatório *n*-dimensional com a totalidade das observações da variável resposta.

$\vec{\mathbf{1}}_n$ o vector de *n* uns.

$\vec{\mathscr{I}}_{A_i}$ a variável indicatriz de pertença ao nível *i* do Factor A.

$\vec{\mathscr{I}}_{B_j}$ a variável indicatriz de pertença ao nível *j* do Factor B.

$\vec{\boldsymbol{\varepsilon}}$ o vector aleatório dos *n* erros aleatórios.

A equação-base em notação vectorial (cont.)

Se se admitissem efeitos para todos os níveis de ambos os factores, temos a equação-base:

$$\vec{\mathbf{Y}} = \mu\vec{\mathbf{1}}_{\mathrm{n}} + \alpha_1\vec{\mathscr{I}}_{A_1} + \alpha_2\vec{\mathscr{I}}_{A_2} + \ldots + \alpha_a\vec{\mathscr{I}}_{A_a} + \beta_1\vec{\mathscr{I}}_{B_1} + \beta_2\vec{\mathscr{I}}_{B_2} + \ldots + \beta_b\vec{\mathscr{I}}_{B_b} + \vec{\boldsymbol{\varepsilon}}$$

A matriz do modelo **X** definida com base nesta equação teria como colunas os vectores $\vec{\mathbf{1}}_{\mathrm{n}}$, $\vec{\mathscr{I}}_{A_1}$, $\vec{\mathscr{I}}_{A_2}$, ..., $\vec{\mathscr{I}}_{A_a}$, $\vec{\mathscr{I}}_{B_1}$, $\vec{\mathscr{I}}_{B_2}$, ..., $\vec{\mathscr{I}}_{B_b}$.

Nessa matriz haveria dependências lineares por duas diferentes razões:

- a soma das indicatrizes do Factor A daria a coluna dos uns, $\vec{\mathbf{1}}_{\mathrm{n}}$;
- a soma das indicatrizes do Factor B daria a coluna dos uns, $\vec{\mathbf{1}}_{\mathrm{n}}$.

Agora, são necessárias duas restrições aos parâmetros, não podendo estimar-se parâmetros α_i e β_j para todos os níveis de cada Factor.

A matriz **X** sem restrições no modelo

$$
\mathbf{X} = \left[\begin{array}{c|ccc|ccc}
1 & 1 & 0 \ \ldots \ 0 & 1 & 0 \ \ldots \ 0 \\
1 & 1 & 0 \ \ldots \ 0 & 1 & 0 \ \ldots \ 0 \\
1 & 1 & 0 \ \ldots \ 0 & 0 & 1 \ \ldots \ 0 \\
\vdots & \vdots & \ddots \ \vdots & \vdots & \ddots \ \vdots \\
1 & 1 & 0 \ \ldots \ 0 & 0 & 0 \ \ldots \ 1 \\
1 & 1 & 0 \ \ldots \ 0 & 0 & 0 \ \ldots \ 1 \\
\hline
1 & 0 & 1 \ \ldots \ 0 & 1 & 0 \ \ldots \ 0 \\
1 & 0 & 1 \ \ldots \ 0 & 1 & 0 \ \ldots \ 0 \\
1 & 0 & 1 \ \ldots \ 0 & 1 & 0 \ \ldots \ 0 \\
\vdots & \vdots & \ddots \ \vdots & \vdots & \ddots \ \vdots \\
1 & 0 & 1 \ \ldots \ 0 & 0 & 0 \ \ldots \ 1 \\
1 & 0 & 1 \ \ldots \ 0 & 0 & 0 \ \ldots \ 1 \\
\hline
\vdots & \vdots & \ddots \ \vdots & \vdots & \ddots \ \vdots \\
\hline
1 & 0 & 0 \ \ldots \ 1 & 1 & 0 \ \ldots \ 0 \\
\vdots & \vdots & \ddots \ \vdots & \vdots & \ddots \ \vdots \\
1 & 0 & 0 \ \ldots \ 1 & 0 & 0 \ \ldots \ 1 \\
1 & 0 & 0 \ \ldots \ 1 & 0 & 0 \ \ldots \ 1
\end{array}\right]
\begin{array}{c|ccc|ccc}
\uparrow & \uparrow & \uparrow & & \uparrow & \uparrow & \uparrow & & \uparrow \\
\vec{1}_n & \vec{\mathscr{I}}_{A_1} & \vec{\mathscr{I}}_{A_2} & \ldots & \vec{\mathscr{I}}_{A_a} & \vec{\mathscr{I}}_{B_1} & \vec{\mathscr{I}}_{B_2} & \ldots & \vec{\mathscr{I}}_{B_b}
\end{array}
$$

A exclusão da coluna $\vec{1}_n$ não resolve o problema.

Equação em notação vectorial, com restrições

Excluímos da equação do modelo as parcelas associadas ao primeiro nível de cada Factor, isto é, impõem-se as duas restrições:

$$\alpha_1 = 0 \qquad \text{e} \qquad \beta_1 = 0 \ ,$$

o que corresponde a excluir as colunas $\vec{\mathscr{I}}_{A_1}$ e $\vec{\mathscr{I}}_{B_1}$ da matriz **X**.

A equação-base do modelo ANOVA a 2 Factores, sem interacção, fica:

$$\vec{\mathbf{Y}} = \mu \vec{\mathbf{1}}_{\mathrm{n}} + \alpha_2 \vec{\mathscr{I}}_{A_2} + \ldots + \alpha_a \vec{\mathscr{I}}_{A_a} + \beta_2 \vec{\mathscr{I}}_{B_2} + \ldots + \beta_b \vec{\mathscr{I}}_{B_b} + \vec{\boldsymbol{\varepsilon}}$$

O parâmetro μ fica o valor esperado das observações na célula $(1,1)$:

$$Y_{11k} = \mu + \varepsilon_{11k} \qquad \Rightarrow \qquad E[Y_{11k}] = \mu = \mu_{11} \ .$$

A matriz do delineamento na ANOVA a 2 Factores (sem interacção), com as restrições $\alpha_1 = 0$ e $\beta_1 = 0$

$$
\mathbf{X} = \left[\begin{array}{c|ccc|ccc}
1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
1 & 0 & \cdots & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & \cdots & 0 & 0 & \cdots & 1 \\
1 & 0 & \cdots & 0 & 0 & \cdots & 1 \\
\hline
1 & 1 & \cdots & 0 & 0 & \cdots & 0 \\
1 & 1 & \cdots & 0 & 0 & \cdots & 0 \\
1 & 1 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & \cdots & 0 & 0 & \cdots & 1 \\
1 & 1 & \cdots & 0 & 0 & \cdots & 1 \\
\hline
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\hline
1 & 0 & \cdots & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & \cdots & 1 & 0 & \cdots & 1 \\
1 & 0 & \cdots & 1 & 0 & \cdots & 1
\end{array}\right]
$$

$\uparrow \vec{1}_n$ $\quad \uparrow \vec{\mathscr{I}}_{A_2}$ $\quad \cdots$ $\quad \uparrow \vec{\mathscr{I}}_{A_a}$ $\quad \uparrow \vec{\mathscr{I}}_{B_2}$ $\quad \cdots$ $\quad \uparrow \vec{\mathscr{I}}_{B_b}$

O modelo ANOVA a dois factores, sem interacção

Juntando os pressupostos necessários à inferência,

Modelo ANOVA a dois factores, sem interacção

Existem n observações, Y_{ijk}, n_{ij} das quais associadas à célula (i,j) $(i=1,...,a;\ j=1,...,b)$. Tem-se:

1. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk}$, $\forall i=1,...,a;\ j=1,...,b;\ k=1,...,n_{ij}$ $\quad (\alpha_1 = 0; \beta_1 = 0)$.
2. $\varepsilon_{ijk} \frown \mathcal{N}(0, \sigma^2), \quad \forall i,j,k$
3. $\{\varepsilon_{ijk}\}_{i,j,k}$ v.a.s independentes.

O modelo tem $a+b-1$ parâmetros desconhecidos:

- o parâmetro μ_{11};
- os $a-1$ acréscimos α_i $(i>1)$; e
- os $b-1$ acréscimos β_j $(j>1)$.

Testando a existência de efeitos

Um teste de ajustamento global do modelo tem como hipótese nula que todos os efeitos, quer do factor A, quer do Factor B são simultaneamente nulos, mas não distingue entre os efeitos de cada factor.

Mais útil será testar separadamente a existência dos efeitos de cada factor. Seria útil dispôr de dois testes, para as hipóteses:

- Teste I: $H_0 : \alpha_i = 0\,, \quad \forall i = 2, ..., a$;
- Teste II: $H_0 : \beta_j = 0\,, \quad \forall j = 2, ..., b$.

Teste aos efeitos do Factor B

O modelo ANOVA a 2 Factores, sem interacção (Acetato 280) tem equação vectorial:

$$\vec{\mathbf{Y}} = \mu\vec{\mathbf{1}}_{\mathrm{n}} + \alpha_2\vec{\mathscr{I}}_{A_2} + \ldots + \alpha_a\vec{\mathscr{I}}_{A_a} + \beta_2\vec{\mathscr{I}}_{B_2} + \ldots + \beta_b\vec{\mathscr{I}}_{B_b} + \vec{\boldsymbol{\varepsilon}}$$

Sendo um Modelo Linear pode-se aplicar a teoria conhecida para este tipo de modelos e testar as hipóteses:

$$H_0 : \beta_j = 0\,, \quad \forall j = 2,\ldots,b \qquad \text{vs.} \qquad H_1 : \exists j \quad \text{tal que } \beta_j \neq 0\ ,$$

através dum teste F parcial comparando o modelo completo

$$(\text{Modelo } M_{A+B}) \qquad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk}\ ,$$

com o submodelo de equação de base

$$(\text{Modelo } M_A) \qquad Y_{ijk} = \mu_{11} + \alpha_i + \varepsilon_{ijk}\ ,$$

que é um modelo ANOVA a 1 Factor (factor A).

A construção do teste aos efeitos do Factor B

Assim,

- Ajusta-se o modelo completo M_{A+B} e o submodelo M_A.
- Obtêm-se as respectivas Somas de Quadrados Residuais, que designamos $SQRE_{A+B}$ e $SQRE_A$.
- Efectua-se o teste F parcial indicado. A estatística de teste é:

$$\text{(Efeitos Factor B)} \quad F = \frac{\overbrace{SQRE_A - SQRE_{A+B}}^{=SQB}}{\frac{SQRE_{A+B}}{n-(a+b-1)}} = \frac{QMB}{QMRE}$$

  definindo $QMB = \frac{SQB}{b-1} = \frac{SQRE_A - SQRE_{A+B}}{b-1}$.

- F tem distribuição $F_{[b-1,\,n-(a+b-1)]}$ sob $H_0 : \beta_j = 0,\ \forall j$.

A construção do teste aos efeitos do Factor A

Consideremos também um teste aos efeitos do Factor A, definido de forma um pouco diferente.

Defina-se:

- $SQA = SQF_A$, a Soma de Quadrados do Factor no Modelo M_A;
- $QMA = \frac{SQA}{a-1}$, o Quadrado Médio do Factor no Modelo M_A;
- $SQRE_{A+B}$ e $QMRE = \frac{SQRE_{A+B}}{n-(a+b-1)}$, como antes.

É possível provar que, caso $\alpha_i = 0$, $\forall i=2,\ldots,a$, a estatística

$$F = \frac{QMA}{QMRE} = \frac{\frac{SQA}{a-1}}{\frac{SQRE_{A+B}}{n-(a+b-1)}}$$

tem distribuição $F_{(a-1,n-(a+b-1))}$.

O Teste F aos efeitos do factor A

Sendo válido o Modelo de ANOVA a dois factores, sem interacção:

Teste F aos efeitos do factor A

Hipóteses: $H_0 : \alpha_i = 0 \quad \forall i=2,...,a$ vs. $H_1 : \exists i=2,..,a \text{ t.q. } \alpha_i \neq 0.$

[A NÃO AFECTA Y] vs. [A AFECTA Y]

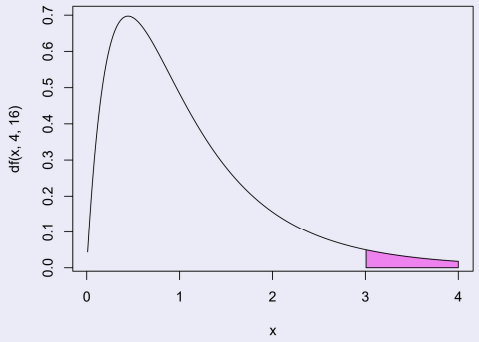
Estatística do Teste: $F = \frac{QMA}{QMRE} \frown F_{(a-1,n-(a+b-1))}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha(a-1,n-(a+b-1))}$



O Teste F aos efeitos do factor B

Sendo válido o Modelo de ANOVA a dois factores, sem interacção:

Teste F aos efeitos do factor B

Hipóteses: $H_0 : \beta_j = 0 \quad \forall j=2,\ldots,b$ vs. $H_1 : \exists j=2,..,b \text{ t.q. } \beta_j \neq 0.$

[B NÃO AFECTA Y] vs. [B AFECTA Y]

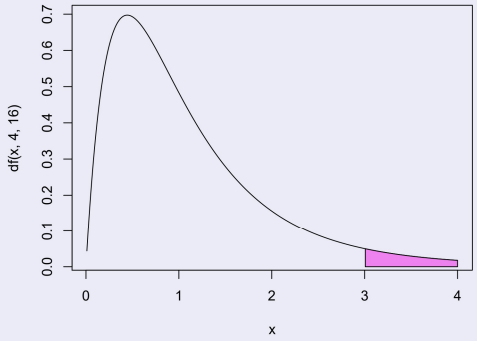
Estatística do Teste: $F = \frac{QMB}{QMRE} \frown F_{(b-1,n-(a+b-1))}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha(b-1,n-(a+b-1))}$



A nova decomposição de SQT

Tendo em conta as Somas de Quadrados antes definidas, tem-se:

$$\begin{aligned} SQB &= SQRE_A - SQRE_{A+B} \\ SQA &= SQF_A = SQT - SQRE_A \end{aligned}$$

Somando estas SQs a $SQRE_{A+B}$, obtém-se:

A decomposição de SQT

$$SQA + SQB + SQRE_{A+B} = SQT$$

que é uma nova decomposição de SQT, em três parcelas, associadas ao facto de haver agora dois factores com efeitos previstos no modelo, mais a variabilidade residual.

Quadro-resumo ANOVA a 2 Factores (sem interacção)

Fonte	g.l.	SQ	QM	f_{calc}
Factor A	$a-1$	$SQA = SQF_A$	$QMA = \frac{SQA}{a-1}$	$\frac{QMA}{QMRE}$
Factor B	$b-1$	$SQB = SQRE_A - SQRE_{A+B}$	$QMB = \frac{SQB}{b-1}$	$\frac{QMB}{QMRE}$
Resíduos	$n-(a+b-1)$	$SQRE = SQRE_{A+B}$	$QMRE = \frac{SQRE}{n-(a+b-1)}$	
Total	$n-1$	$SQT = (n-1)\,s_y^2$	–	–

ANOVA a dois Factores, sem interacção no R

Para efectuar uma ANOVA a dois Factores (sem interacção) no R, convém organizar os dados numa `data.frame` com três colunas:

1. uma para os valores (numéricos) da variável resposta;
2. outra para o factor A (com a indicação dos seus níveis);
3. outra para o factor B (com a indicação dos seus níveis).

As fórmulas utilizadas no R para indicar uma ANOVA a dois Factores, sem interacção, são semelhantes às usadas na Regressão Linear com dois preditores, devendo o nome dos dois factores ser separado pelo símbolo $+$:

```
y  ~  fA + fB
```

Um exemplo clássico: os rendimentos de cevada

O rendimento de $a = 5$ variedades de cevada (*manchuria, svansota, velvet, trebi* e *peatland*) foi registado em $b = 6$ diferentes localidades [a]. Em cada localidade foi semeada (com casualização) uma parcela com cada variedade ($n = 30$).

```
> summary(aov(Y1 ~ Var + Loc, data=immer))
            Df  Sum Sq Mean Sq F value    Pr(>F)
Var          4  2756.6   689.2  4.2309   0.01214 *
Loc          5 17829.8  3566.0 21.8923 1.751e-07 ***
Residuals   20  3257.7   162.9
```

Há indicação de efeitos significativos (ao nível $\alpha = 0.05$) entre variedades e muito significativos entre localidades. Num modelo ignorando os efeitos de localidades, desaparecia a significância dos efeitos de variedade:

```
> summary(aov(Y1 ~ Var, data=immer))
            Df  Sum Sq Mean Sq F value Pr(>F)
Var          4  2756.6   689.2   0.817 0.5264
Residuals   25 21087.6   843.5
```

[a] Dados em Immer, Hayes e LeRoy Powers, Statistical adaptation of barley varietal adaptation, Journal of the American Society for Agronomy, 26, 403-419, 1934.

Trocando a ordem dos factores

Atenção: A forma como foram definidas as Somas de Quadrados de cada factor é diferente: $SQB = SQRE_A - SQRE_{A+B}$ e $SQA = SQF_A$.

A troca do papel dos factores A e B produz resultados diferentes em delineamentos não equilibrados. Designando por M_B o modelo ANOVA a um factor, mas apenas com o factor que temos chamado B, tem-se:

$$\begin{aligned} SQB &= SQF_B = SQT - SQRE_B \\ SQA &= SQRE_B - SQRE_{A+B} \ . \end{aligned}$$

Continua a ser verdade que SQT se pode decompor na forma

$$SQT = SQA + SQB + SQRE_{A+B} \ .$$

Justificam-se testes análogos aos dos acetatos 285 e 286.
Mas as duas formas alternativas de definir SQA e SQB apenas produzem resultados iguais no caso de delineamentos equilibrados, pelo que só nesse caso a ordem dos factores é arbitrária. (Ver também o Ex. ANOVA 9)

As várias médias amostrais

Sejam, num delineamento equilibrado:

$\overline{Y}_{i..}$ a média amostral das $b\,n_c$ observações do nível i do Factor A, $\overline{Y}_{i..} = \frac{1}{b\,n_c} \sum_{j=1}^{b} \sum_{k=1}^{n_c} Y_{ijk}$

$\overline{Y}_{.j.}$ a média amostral das $a\,n_c$ observações do nível j do Factor B, $\overline{Y}_{.j.} = \frac{1}{a\,n_c} \sum_{i=1}^{a} \sum_{k=1}^{n_c} Y_{ijk}$

$\overline{Y}_{...}$ a média amostral da totalidade das $n = a\,b\,n_c$ observações, $\overline{Y}_{...} = \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_c} Y_{ijk}$.

SQA e SQB em delineamentos equilibrados

Num delineamento equilibrado, SQA é igual à Soma de Quadrados do Factor (SQF_A) do Modelo M_A, apenas com o Factor A (acetato 284).

Nesse modelo, os valores ajustados são $\hat{Y}_{ijk} = \overline{Y}_{i..}$ (acetato 240). Assim, num delineamento equilibrado, tem-se:

$$SQF_A = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_c}(\underbrace{\hat{Y}_{ijk}}_{=\overline{Y}_{i..}} - \overline{Y}_{...})^2 = b\,n_c \cdot \sum_{i=1}^{a}(\overline{Y}_{i..} - \overline{Y}_{...})^2 = SQA\ .$$

Da mesma forma, num delineamento equilibrado, SQB é a Soma de Quadrados do Factor (SQF_B) do Modelo M_B, apenas com o Factor B. Nesse modelo, os valores ajustados são $\hat{Y}_{ijk} = \overline{Y}_{.j.}$, logo:

$$SQF_B = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_c}(\underbrace{\hat{Y}_{ijk}}_{=\overline{Y}_{.j.}} - \overline{Y}_{...})^2 = a\,n_c \cdot \sum_{j=1}^{b}(\overline{Y}_{.j.} - \overline{Y}_{...})^2 = SQB\ .$$

Fórmulas para delineamentos equilibrados (cont.)

Se o delineamento é equilibrado, ou seja, $n_{ij} = n_c \ , \ \forall i,j$, tem-se:

- $\hat{\mu}_{11} = \overline{Y}_{1..} + \overline{Y}_{.1.} - \overline{Y}_{...}$
- $\hat{\alpha}_i = \overline{Y}_{i..} - \overline{Y}_{1..}$
- $\hat{\beta}_j = \overline{Y}_{.j.} - \overline{Y}_{.1.}$

Tendo em conta a equação base do Modelo, os valores ajustados de cada observação dependem apenas das médias dos respectivos níveis em cada factor e da média geral de todas as observações:

$$\hat{Y}_{ijk} = \hat{\mu}_{11} + \hat{\alpha}_i + \hat{\beta}_j = \overline{Y}_{i..} + \overline{Y}_{.j.} - \overline{Y}_{...} \quad , \quad \forall i,j,k$$

Aviso: Ao contrário do que sucede na ANOVA a um factor, os valores ajustados $\hat{Y}_{ijk}$ não são a média das observações de Y na célula (i,j).

O quadro-resumo da ANOVA a 2 Factores (sem interacção; delineamento equilibrado)

Fonte	g.l.	SQ	QM	f_{calc}
Factor A	$a-1$	$SQA = b\,n_c \cdot \sum_{i=1}^{a} (\overline{y}_{i..} - \overline{y}_{...})^2$	$QMA = \frac{SQA}{a-1}$	$\frac{QMA}{QMRE}$
Factor B	$b-1$	$SQB = a\,n_c \cdot \sum_{j=1}^{b} (\overline{y}_{.j.} - \overline{y}_{...})^2$	$QMB = \frac{SQB}{b-1}$	$\frac{QMB}{QMRE}$
Resíduos	$n-(a+b-1)$	$SQRE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_c} [y_{ijk} - (\overline{y}_{i..} + \overline{y}_{.j.} - \overline{y}_{...})]^2$	$QMRE = \frac{SQRE}{n-(a+b-1)}$	
Total	$n-1$	$SQT = (n-1)\,s_y^2$	–	–

A interpretação dos parâmetros

A interpretação do significado dos parâmetros do modelo depende da convenção usada para resolver o problema da multicolinearidade das colunas da matriz **X**.

Vejamos a interpretação dos parâmetros resultante da convenção $\alpha_1 = \beta_1 = 0$.

Uma observação de Y efectuada na célula $(1,1)$, correspondente ao cruzamento do primeiro nível de cada factor, será da forma:

$$Y_{11k} = \mu_{11} + \underbrace{\alpha_1}_{=0} + \underbrace{\beta_1}_{=0} + \varepsilon_{11k} \qquad \Longrightarrow \qquad E[Y_{11k}] = \mu_{11}$$

O parâmetro μ_{11} corresponde ao valor esperado da variável resposta Y na célula cujas indicatrizes foram excluídas da matriz do delineamento.

A interpretação dos parâmetros α_i

Uma observação de Y efectuada na célula $(i,1)$, com $i>1$ (cruzamento dum nível do factor A diferente do primeiro, com o primeiro nível do Factor B) é da forma:

$$Y_{i1k} = \mu_{11} + \alpha_i + \underbrace{\beta_1}_{=0} + \varepsilon_{i1k} \qquad \Longrightarrow \qquad \mu_{i1} = E[Y_{i1k}] = \mu_{11} + \alpha_i$$

O parâmetro $\alpha_i = \mu_{i1} - \mu_{11}$ corresponde ao acréscimo no valor esperado da variável resposta Y associado a observações do nível $i>1$ do Factor A (relativamente às observações do primeiro nível do Factor A), quando $j=1$. Designa-se o efeito do nível i do factor A.

Interpretação dos parâmetros α_i

Tabela com médias populacionais de célula (situação experimental):

		Factor B				
	Níveis	B_1	B_2	B_3	$\ldots$	B_b
FACTOR A	A_1	μ_{11}	μ_{12}	μ_{13}	$\ldots$	μ_{1b}
	A_2	$\mu_{21} = \mu_{11} + \alpha_2$	μ_{22}	μ_{23}	$\ldots$	μ_{2b}
	A_3	$\mu_{31} = \mu_{11} + \alpha_3$	μ_{32}	μ_{33}	$\ldots$	μ_{3b}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	A_a	$\mu_{a1} = \mu_{11} + \alpha_a$	μ_{a2}	μ_{a3}	$\ldots$	μ_{ab}

A interpretação dos parâmetros β_j

Uma observação de Y efectuada na célula $(1,j)$, com $j > 1$ (cruzamento do primeiro nível do factor A com um nível do Factor B diferente do primeiro) é da forma:

$$Y_{1jk} = \mu_{11} + \underbrace{\alpha_1}_{=0} + \beta_j + \varepsilon_{1jk} \qquad \Longrightarrow \qquad \mu_{1j} = E[Y_{1jk}] = \mu_{11} + \beta_j$$

O parâmetro $\beta_j = \mu_{1j} - \mu_{11}$ corresponde ao acréscimo no valor esperado da variável resposta Y associado a observações do nível j do Factor B (relativamente às observações do primeiro nível do Factor B), quando $i = 1$. Designa-se o efeito do nível j do factor B.

Interpretação dos parâmetros β_j

Tabela com médias populacionais de célula (situação experimental):

		Factor B				
	Níveis	B_1	B_2	B_3	...	B_b
Factor A	A_1	μ_{11}	$\mu_{12} = \mu_{11} + \beta_2$	$\mu_{13} = \mu_{11} + \beta_3$	...	$\mu_{1b} = \mu_{11} + \beta_b$
	A_2	μ_{21}	μ_{22}	μ_{23}	...	μ_{2b}
	A_3	μ_{31}	μ_{32}	μ_{33}	...	μ_{3b}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	A_a	μ_{a1}	μ_{a2}	μ_{a3}	...	μ_{ab}

Observações de Y no caso geral

Mas este modelo é pouco flexível: não existem mais parâmetros e os valores esperados nas restantes células já estão fixados.

Para observações de Y efectuadas numa célula genérica (i,j), com $i > 1$ e $j > 1$, tem-se:

$$Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk} \qquad \Longrightarrow \qquad \mu_{ij} = E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_j .$$

Todas as parcelas destes valores esperados de Y já foram usados. Não há flexibilidade para descrever as médias de células com $i > 1$ e $j > 1$.

Um modelo sem efeitos de interacção é utilizado sobretudo quando existe uma única observação em cada célula, i.e., $n_{ij} = 1, \forall\, i,j$.

Modelos com interacção

Um modelo ANOVA a 2 Factores, sem interacção, foi considerado para um delineamento factorial, isto é, em que se cruzam todos os níveis de um e outro factor. Mas trata-se dum modelo pouco flexível.

Na presença de repetições nas células, a forma mais natural de modelar um delineamento com dois factores é a de prever a existência de um terceiro tipo de efeitos: os efeitos de interacção.

A ideia é incorporar na equação base do modelo para Y_{ijk} uma parcela $(\alpha\beta)_{ij}$ que permita que em cada célula haja um efeito específico associado à combinação dos níveis *i* do Factor A e *j* do Factor B:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \ .$$

Os valores esperados de Y_{ijk} (modelo com interacção)

Vamos admitir as seguintes restrições aos parâmetros:

$$\alpha_1 = 0 \quad ; \quad \beta_1 = 0 \quad ; \quad (\alpha\beta)_{1j} = 0\ ,\ \forall j \quad ; \quad (\alpha\beta)_{i1} = 0\ ,\ \forall i.$$

Tem-se, a partir da equação $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$:

- Para a primeira célula ($i = j = 1$): $\mu_{11} = E[Y_{11k}] = \mu$.
- Nas restantes células $(1, j)$ do primeiro nível do Factor A:
  $\mu_{1j} = E[Y_{1jk}] = \mu_{11} + \beta_j$.
- Nas restantes células $(i, 1)$ do primeiro nível do Factor B:
  $\mu_{i1} = E[Y_{i1k}] = \mu_{11} + \alpha_i$.
- Nas células genéricas (i, j), com $i > 1$ e $j > 1$,
  $\mu_{ij} = E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$.

Os efeitos α_i e β_j designam-se efeitos principais de cada Factor.

Os valores esperados de Y_{ijk} (modelo com interacção)

Efeito das restrições $\alpha_1 = 0$; $\beta_1 = 0$; $(\alpha\beta)_{ij} = 0$ se $i = 1$ ou $j = 1$:

		Factor B				
	Níveis	B_1	B_2	B_3	...	B_b
FACTOR A	A_1	× × ×	× × ×	× × ×	...	× × ×
	A_2	× × ×	× × ×	× × ×	...	× × ×
	A_3	× × ×	× × ×	× × ×	...	× × ×
	⋮	⋮	⋮	⋮	⋱	⋮
	A_a	× × ×	× × ×	× × ×	...	× × ×

As observações que não estão associadas a A_1 (primeira linha) têm efeitos α_i.

As observações que não estão associadas a B_1 (primeira coluna) têm efeitos β_j.

As observações que não são da primeira coluna nem da primeira linha têm efeitos de interacção $(\alpha\beta)_{ij}$.

O modelo ANOVA a dois factores, com interacção

Juntando os pressupostos necessários à inferência,

Modelo ANOVA a dois factores, com interacção (Modelo M_{A*B})

Existem n observações, Y_{ijk}, n_{ij} das quais associadas à célula (i,j) $(i = 1,...,a; j = 1,...,b)$. Tem-se:

1. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, $\quad \forall i{=}1,...,a \ ; \ j{=}1,...,b \ ; \ k{=}1,...,n_{ij}$
   $(\alpha_1{=}0 \ ; \ \beta_1{=}0 \ ; \ (\alpha\beta)_{ij}{=}0$, se $i{=}1$ e/ou $j{=}1)$.
2. $\varepsilon_{ijk} \frown \mathcal{N}(0, \sigma^2)$
3. $\{\varepsilon_{ijk}\}_{i,j,k}$ v.a.s independentes.

O modelo tem ab parâmetros desconhecidos:

- a 1 média da célula de referência, μ_{11};
- os $a-1$ acréscimos α_i $(i > 1)$;
- os $b-1$ acréscimos β_j $(j > 1)$; e
- os $(a-1)(b-1)$ efeitos de interacção $(\alpha\beta)_{ij}$, para $i > 1, j > 1$.

Variáveis indicatrizes de célula

A versão vectorial da equação do modelo com interacção associa os novos efeitos $(\alpha\beta)_{ij}$ a variáveis indicatrizes das respectivas células.

A equação-base do modelo ANOVA a 2 Factores, com interacção, é:

$$\begin{aligned}\vec{\mathbf{Y}} \;=\; & \mu\vec{\mathbf{1}}_{\mathrm{n}} + \alpha_2\vec{\mathscr{I}}_{A_2} + \ldots + \alpha_a\vec{\mathscr{I}}_{A_a} + \beta_2\vec{\mathscr{I}}_{B_2} + \ldots + \beta_b\vec{\mathscr{I}}_{B_b} + \\ & + (\alpha\beta)_{22}\vec{\mathscr{I}}_{A_2:B_2} + (\alpha\beta)_{23}\vec{\mathscr{I}}_{A_2:B_3} + \ldots + (\alpha\beta)_{ab}\vec{\mathscr{I}}_{A_a:B_b} + \vec{\boldsymbol{\varepsilon}}\end{aligned}$$

onde $\vec{\mathscr{I}}_{A_i:B_j}$ representa a variável indicatriz da célula correspondente ao nível *i* do Factor A e nível *j* do factor B.

Este modelo com *ab* parâmetros é designado modelo M_{A*B}

Modelo ANOVA a 2 factores, com interacção (cont.)

A matriz **X** do delineamento é agora constituída por ab colunas:

- uma coluna de uns, $\vec{\mathbf{1}}_n$, associada ao parâmetro μ_{11}.
- $a-1$ colunas de indicatrizes de nível do factor A, $\vec{\mathscr{I}}_{A_i}$, $(i>1)$, associadas aos parâmetros α_i.
- $b-1$ colunas de indicatrizes de nível do factor B, $\vec{\mathscr{I}}_{B_j}$, $(j>1)$, associadas aos parâmetros β_j.
- $(a-1)(b-1)$ colunas de indicatrizes de célula, $\vec{\mathscr{I}}_{A_i:B_j}$, $(i,j>1)$, associadas aos efeitos de interacção $(\alpha\beta)_{ij}$.

Como em modelos anteriores, $\hat{\vec{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$, sendo **H** a matriz que projecta ortogonalmente sobre o espaço $\mathscr{C}(\mathbf{X})$ gerado pelas colunas desta matriz **X**.

E também, $SQRE_{A*B} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(Y_{ijk}-\hat{Y}_{ijk})^2$.

Os três testes ANOVA

Neste delineamento, desejamos fazer um teste à existência de cada um dos três tipos de efeitos:

- Teste I: $H_0 : (\alpha\beta)_{ij} = 0\,, \quad \forall i = 2,...,a\ ,\ \forall j = 2,...,b$;
- Teste II: $H_0 : \alpha_i = 0\,, \quad \forall i = 2,...,a$; e
- Teste III: $H_0 : \beta_j = 0\,, \quad \forall j = 2,...,b$.

As estatísticas de teste para cada um destes três testes obtêm-se a partir da decomposição da Soma de Quadrados Total (ou seja, da *análise da variancia*) em parcelas convenientes.

Testando efeitos de interacção

Para testar a existência de efeitos de interacção,

$$H_0 : (\alpha\beta)_{ij} = 0\,, \quad \forall i = 2,...,a\ ,\ \forall j = 2,...,b\ ,$$

pode efectuar-se um teste F parcial comparando o modelo

$$(\text{Modelo } M_{A*B}) \qquad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}\ ,$$

com o submodelo sem efeitos de interacção

$$(\text{Modelo } M_{A+B}) \qquad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk}\ ,$$

Designa-se Soma de Quadrados associada à interacção à diferença

$$SQAB = SQRE_{A+B} - SQRE_{A*B}$$

Testando os efeitos principais de cada Factor

Para testar os efeitos principais dos Factor B ($H_0 : \beta_j = 0\,,\ \forall j = 2, ..., b$) e do Factor A ($H_0 : \alpha_i = 0\,,\ \forall i = 2, ..., a$) pode partir-se dos modelos

$$\begin{aligned} (\text{Modelo } M_{A+B}) \qquad Y_{ijk} &= \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk} \\ (\text{Modelo } M_A) \qquad Y_{ijk} &= \mu_{11} + \alpha_i + \varepsilon_{ijk}\ , \end{aligned}$$

Defina-se:

$$\begin{aligned} SQB &= SQRE_A - SQRE_{A+B} \\ SQA &= SQF_A = SQT - SQRE_A \end{aligned}$$

Nota: Estas duas Somas de Quadrados definem-se da mesma forma que no modelo sem efeitos de interacção.

A decomposição de SQT

Definimos :

$$\begin{aligned} SQAB &= SQRE_{A+B} - SQRE_{A*B} \\ SQB &= SQRE_A - SQRE_{A+B} \\ SQA &= SQF_A = SQT - SQRE_A \end{aligned}$$

Somando estas Somas de Quadrados a $SQRE_{A*B}$, obtém-se:

$$SQT = SQRE_{A*B} + SQAB + SQA + SQB$$

Esta decomposição de SQT gera as quantidades nas quais se baseiam as estatísticas dos três testes associados ao Modelo M_{A*B}.

O quadro-resumo

Com base na decomposição do acetato 311 podemos construir o quadro resumo da ANOVA a 2 Factores, com interacção.

Fonte	g.l.	SQ	QM	f_{calc}
Factor A	$a-1$	SQA	$QMA = \frac{SQA}{a-1}$	$\frac{QMA}{QMRE}$
Factor B	$b-1$	SQB	$QMB = \frac{SQB}{b-1}$	$\frac{QMB}{QMRE}$
Interacção	$(a-1)(b-1)$	$SQAB$	$QMAB = \frac{SQAB}{(a-1)(b-1)}$	$\frac{QMAB}{QMRE}$
Resíduos	$n-ab$	$SQRE$	$QMRE = \frac{SQRE}{n-ab}$	
Total	$n-1$	$SQT = (n-1)\,s_y^2$	–	–

Os graus de liberdade de cada tipo de efeito são o número de parâmetros desse tipo que sobram após a imposição das restrições.

Como em qualquer modelo linear, os graus de liberdade residuais são o número de observações (n) menos o número de parâmetros do modelo (ab).

O Teste F aos efeitos de interacção

Sendo válido o Modelo ANOVA a dois factores, com interacção:

Teste F aos efeitos de interacção

Hipóteses: $H_0 : (\alpha\beta)_{ij} = 0 \quad \forall i,j$ vs. $H_1 : \exists i,j \text{ t.q. } (\alpha\beta)_{ij} \neq 0$.

[NÃO HÁ INTERACÇÃO] vs. [HÁ INTERACÇÃO]

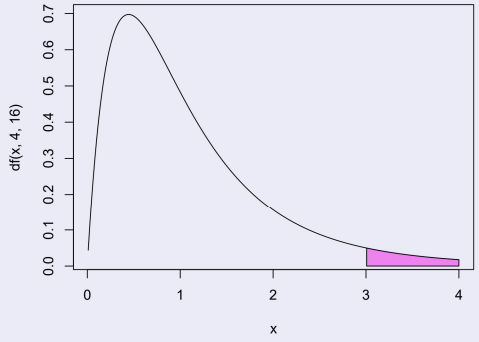
Estatística do Teste: $F = \frac{QMAB}{QMRE} \frown F_{((a-1)(b-1),n-ab)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha((a-1)(b-1),n-ab)}$



O Teste F aos efeitos principais do factor A

Sendo válido o Modelo ANOVA a 2 factores com interacção tem-se:

Teste F aos efeitos principais do factor A

Hipóteses: $H_0 : \alpha_i = 0 \quad \forall i=2,\ldots,a$ vs. $H_1 : \exists i=2,..,a \text{ t.q. } \alpha_i \neq 0.$

[$\nexists$ EFEITOS DE A] vs. [$\exists$ EFEITOS DE A]

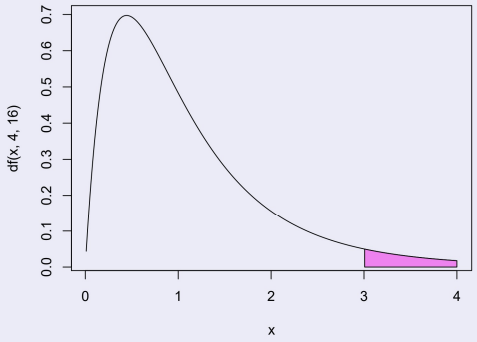
Estatística do Teste: $F = \frac{QMA}{QMRE} \frown F_{(a-1,n-ab)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha(a-1,n-ab)}$



O Teste F aos efeitos principais do factor B

Sendo válido o Modelo ANOVA a 2 factores com interacção tem-se:

Teste F aos efeitos principais do factor B

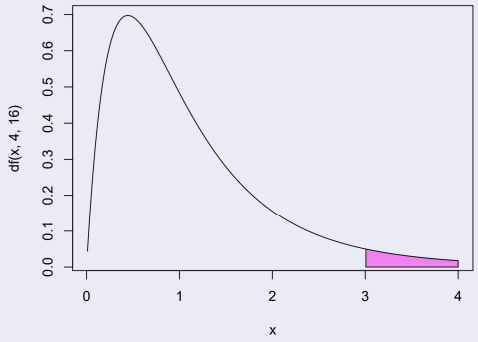
Hipóteses: $H_0 : \beta_j = 0 \quad \forall j=2,\ldots,b$ vs. $H_1 : \exists j=2,..,b \text{ t.q. } \beta_j \neq 0$.
[$\nexists$ EFEITOS DE B] vs. [$\exists$ EFEITOS DE B]

Estatística do Teste: $F = \frac{QMB}{QMRE} \frown F_{(b-1,n-ab)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se
$F_{calc} > f_{\alpha(b-1,n-ab)}$



ANOVA a dois Factores, com interacção no R

Para efectuar uma ANOVA a dois Factores, com interacção, no R, organizam-se os dados de forma igual à usada para o modelo sem interacção: uma `data.frame` com três colunas:

1. uma para a variável resposta;
2. outra para o factor A;
3. outra para o factor B.

As fórmulas utilizadas no R para indicar uma ANOVA a dois Factores, com interacção, recorrem ao símbolo `*`:

```
y  ~  fA * fB
```

sendo `y` o nome da variável resposta e `fA` e `fB` os nomes dos factores.

Estimação da interacção necessita de repetições

Para se poder estudar efeitos de interacção, é necessário que haja repetições nas células.

Os graus de liberdade do *SQRE* neste modelo são $n - ab$. Se houver uma única observação em cada célula, tem-se $n = ab$, ou seja, tantos parâmetros quantas as observações existentes. Nesse caso, nem sequer será possível definir o Quadrado Médio Residual, *QMRE*.

Num delineamento com uma única observação por célula é obrigatório optar por um modelo sem interacção.

Havendo repetições, é mais natural considerar um modelo com interacção e deixar que a conclusão sobre a existência, ou não, desse tipo de efeitos resulte do estudo do modelo.

Não constando do modelo, eventuais efeitos de interacção irão inflacionar a variabilidade residual, não explicada pelo modelo.

Valores ajustados de Y no modelo com interacção

Às médias já definidas no estudo do modelo a dois Factores, sem efeitos de interacção, (acetato 292):

$\overline{Y}_{i..}$ - nível i do Factor A;

$\overline{Y}_{.j.}$ - nível j do Factor B;

$\overline{Y}_{...}$ - global;

acrescentam-se agora as médias de cada célula:

$$\overline{Y}_{ij.} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk} \ .$$

Os valores ajustados $\hat{Y}_{ijk}$ são iguais para todas as observações numa mesma célula, e são dados pela média amostral da célula:

$$\hat{Y}_{ijk} = \overline{Y}_{ij.} \ .$$

Estimadores de parâmetros

Os estimadores dos parâmetros num modelo ANOVA a 2 Factores, com interacção, são dadas pelas quantidades amostrais correspondentes às definições populacionais de cada parâmetro (ver acetato 303):

- $\mu = \mu_{11} \quad \Rightarrow \quad \hat{\mu} = \hat{\mu}_{11} = \overline{Y}_{11\cdot}$
- $\alpha_i = \mu_{i1} - \mu_{11} \quad \Rightarrow \quad \hat{\alpha}_i = \overline{Y}_{i1\cdot} - \overline{Y}_{11\cdot} \qquad (i > 1)$
- $\beta_j = \mu_{1j} - \mu_{11} \quad \Rightarrow \quad \hat{\beta}_j = \overline{Y}_{1j\cdot} - \overline{Y}_{11\cdot} \qquad (j > 1)$
- $(\alpha\beta)_{ij} = \mu_{ij} - \cancel{\mu_{11}} - \underbrace{\alpha_i}_{=\mu_{i1}-\cancel{\mu_{11}}} - \underbrace{\beta_j}_{=\mu_{1j}-\mu_{11}} = \mu_{ij} + \mu_{11} - \mu_{i1} - \mu_{1j}$

  $\Rightarrow \quad \widehat{(\alpha\beta)}_{ij} = (\overline{Y}_{ij\cdot} + \overline{Y}_{11\cdot}) - (\overline{Y}_{i1\cdot} + \overline{Y}_{1j\cdot}) \qquad (i,j > 1)$

Intervalos de confiança ou testes de hipóteses para qualquer parâmetro individual, ou combinações lineares desses parâmetros, podem ser efectuados utilizando a teoria geral do Modelo Linear.

Soma de Quadrados Residual

Como os valores ajustados correspondem às medias amostrais da célula onde se efectuaram as observações, $\hat{Y}_{ijk} = \overline{Y}_{ij.}$, tem-se:

$$SQRE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} (Y_{ijk} - \hat{Y}_{ijk})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} (Y_{ijk} - \overline{Y}_{ij.})^2$$

$$\Leftrightarrow \quad SQRE = \sum_{i=1}^{a}\sum_{j=1}^{b} (n_{ij} - 1) S_{ij}^2 \ ,$$

sendo S_{ij}^2 a variância amostral das observações de Y na célula (i,j).

Num delineamento equilibrado, tem-se $n = n_c ab$, e o Quadrado Médio Residual será a média simples das variâncias amostrais de célula, S_{ij}^2:

$$QMRE = \frac{SQRE}{n - ab} = \frac{\cancel{n_c - 1}}{ab(\cancel{n_c - 1})} \sum_{i=1}^{a}\sum_{j=1}^{b} S_{ij}^2 = \frac{1}{ab} \sum_{i=1}^{a}\sum_{j=1}^{b} S_{ij}^2 \ .$$

Outras SQs para delineamentos equilibrados

Para delineamentos equilibrados (com n_c observações por célula) é possível obter igualmente fórmulas simples para as Somas de Quadrados associadas aos efeitos principais de cada factor.

Estas fórmulas correspondem (tal como no modelo sem efeitos de interacção) às Somas de Quadrados associadas a cada factor, caso se ajustasse (aos mesmos dados) um modelo ANOVA apenas com esse factor:

$$SQA = bn_c \sum_{i=1}^{a} (\overline{Y}_{i..} - \overline{Y}_{...})^2$$

$$SQB = an_c \sum_{j=1}^{b} (\overline{Y}_{.j.} - \overline{Y}_{...})^2$$

Um exemplo:

Dietas de leitões

Variável resposta: Coeficiente de Utilização Digestiva para a celulose (CEL).
Factor A: Fibra (a=2 tipos de fibra).
Factor B: Enzima (b=2 níveis – com e sem enzima na dieta).
Nas ab=4 situações experimentais há $n_{ij} = 12$ repetições (delineamento equilibrado).

```
> leitoes.aov <- aov(CEL ~ Fibra*Enzima , data=leitoes)
> summary(leitoes.aov)
             Df Sum Sq Mean Sq F value  Pr(>F)
Fibra         1 0.0239 0.02385   1.450 0.23500
Enzima        1 0.1376 0.13760   8.364 0.00593 **
Fibra:Enzima  1 0.0257 0.02567   1.560 0.21824
Residuals    44 0.7239 0.01645
```

Neste exemplo, apenas a adição de enzima tem efeito significativo sobre o coeficiente de utilização digestiva.

Exemplo

Dietas de leitões

Como $a=b=2$, há apenas um efeito de cada tipo:

$$\vec{\mathbf{Y}} = \mu\vec{\mathbf{1}}_{\mathrm{n}} + \alpha_2\vec{\mathscr{I}}_{A_2} + \beta_2\vec{\mathscr{I}}_{B_2} + (\alpha\beta)_{22}\vec{\mathscr{I}}_{A_2:B_2} + \vec{\boldsymbol{\varepsilon}}$$

É fácil sintetizar as conclusões:

Teste I:	$H_0 : \alpha_2 = 0$	$p\text{-value} = 0.23500 \Rightarrow$	Não rejeitar $H_0 : \alpha_2 = 0$
Teste II:	$H_0 : \beta_2 = 0$	$p\text{-value} = 0.00593 \Rightarrow$	Optar por $H_1 : \beta_2 \neq 0$
Teste III:	$H_0 : (\alpha\beta)_{2,2} = 0$	$p\text{-value} = 0.21824 \Rightarrow$	Não rejeitar $H_0 : (\alpha\beta)_{2,2} = 0$

		Enzima	
		sem	com
Fibra	1	μ_{11}	$\mu_{12} = \mu_{11} + \beta_2$
	2	$\mu_{21} = \mu_{11} + \alpha_2$	$\mu_{22} = \mu_{11} + \alpha_2 + \beta_2 + (\alpha\beta)_{2,2}$

Comparações múltiplas de médias de células

Havendo *ab* células, a comparação das médias de cada par de células envolve $\binom{ab}{2}$ comparações.

O número potencialmente grande de comparações possíveis entre médias de célula aconselha a utilização de métodos de comparação múltipla, que permitam controlar globalmente o nível de significância do conjunto de testes de hipóteses (ou grau de confiança do conjunto de intervalos de confiança).

O mais utilizado dos métodos de comparação múltipla está associado ao nome de Tukey. Foi já introduzido no estudo de delineamentos a 1 Factor. Adapta-se facilmente à comparação múltipla de médias de células.

O Teste de Tukey

Teste de Tukey para médias de células

Admite-se que o delineamento é equilibrado, com $n_c > 1$ repetiçoes em todas as ab células.

Rejeita-se a igualdade das médias das células (i,j) e (i',j'), a favor da hipótese $\mu_{ij} \neq \mu_{i'j'}$, se

$$|\overline{Y}_{ij\cdot} - \overline{Y}_{i'j'\cdot}| \quad > \quad q_{\alpha\,(ab,n-ab)} \cdot \sqrt{\frac{QMRE}{n_c}} \ ,$$

sendo $q_{\alpha\,(ab,n-ab)}$ o valor que deixa à direita uma região de probabilidade α numa distribuição de Tukey com parâmetros $k = ab$ (o número total de médias de célula) e $\nu = n - ab$ (os graus de liberdade associados ao $QMRE$).

Intervalos de Confiança para $\mu_{ij} - \mu_{i'j'}$

Intervalos de Confiança de Tukey

Com grau de confiança global $(1-\alpha)\times 100\%$, todas as diferenças de médias de pares de células, $\mu_{ij} - \mu_{i'j'}$, estão em intervalos da forma:

$$\left] \left(\overline{y}_{ij\cdot} - \overline{y}_{i'j'\cdot}\right) - q_{\alpha(ab,n-ab)}\sqrt{\frac{QMRE}{n_c}} \quad , \quad \left(\overline{y}_{ij\cdot} - \overline{y}_{i'j'\cdot}\right) + q_{\alpha(ab,n-ab)}\sqrt{\frac{QMRE}{n_c}} \right[$$

Conclui-se que $\mu_{ij} \neq \mu_{i'j'}$ se o intervalo correspondente a este par de células não contém o valor zero.

Tukey no R

A obtenção dos Intervalos de Confiança de Tukey no R, para a diferença da média de células, no caso de um delineamento a dois Factores, é análogo ao caso de um único factor:

```
> TukeyHSD(aov(y ~ fA * fB, data=dados))
```

O R produz também intervalos de confiança para as médias de nível de cada Factor isoladamente. Também pode ser usada a função HSD.test da library(agricolae): > HSD.test(leitoes.aov, c("fA","fB"), console=TRUE)

É possível representar graficamente estes Intervalos de Confiança encaixando o comando anterior na função `plot`.

Visualização gráfica de efeitos de interacção

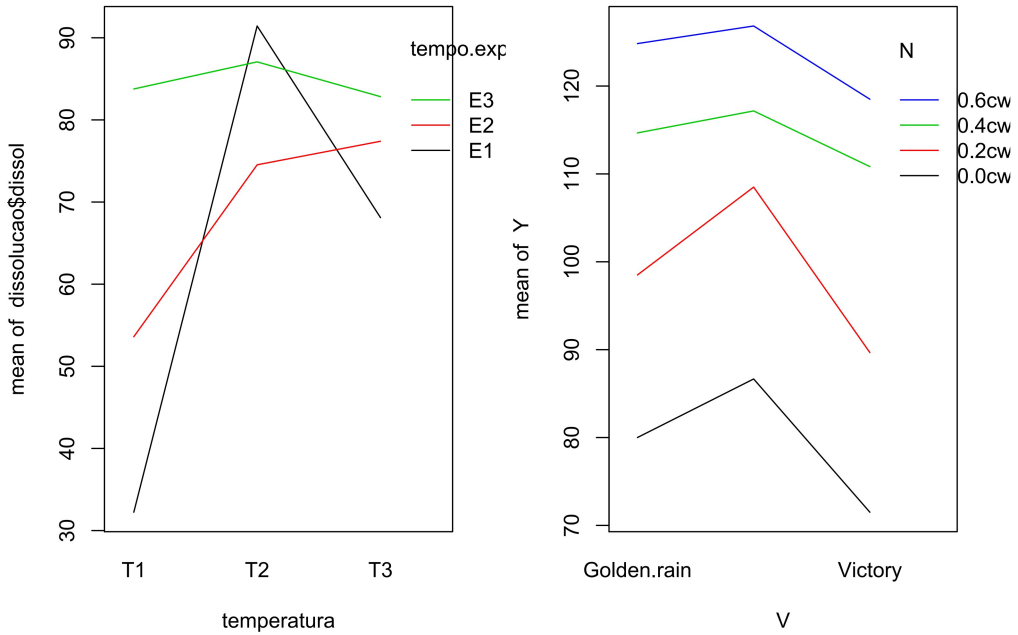
A existência de efeitos de interacção em delineamentos factoriais a dois factores transparece em gráficos onde:

- O eixo horizontal é associado aos níveis de um factor (e.g., *fA*);
- no eixo vertical são indicados os valores médios da variável resposta *Y* em cada célula;
- para cada célula, indica-se um ponto cujas coordenadas são determinadas pelo nível do primeiro factor e respectiva média de célula da variável resposta;
- unem-se com segmentos de recta os pontos correspondentes a um mesmo nível do segundo factor (e.g., *fB*).

A cada problema correspondem sempre dois possíveis gráficos de interacção, pois é arbitrária a escolha de qual o factor associado ao eixo horizontal, e qual o que define os pontos a serem unidos.

Como ler os gráficos de interacção

Havendo interacção, as linhas estarão longe de qualquer paralelismo (exemplo à esquerda). A inexistência de interacção significativa produz linhas aproximadamente “paralelas” (exemplo à direita).



A confirmação da significância dos efeitos de interacção exige que se efectue o respectivo teste *F*.

Análise dos Resíduos

A validade dos pressupostos do Modelo relativos aos erros aleatórios pode ser estudada de forma análoga ao que foi visto para um delineamento a 1 Factor.

Os resíduos relativos a uma mesma célula aparecem em *ab* colunas verticais num gráfico de E_{ijk} vs. $\hat{Y}_{ijk}$.

A hipótese de heterogeneidade de variâncias entre diferentes células pode ser testada recorrendo a testes de hipóteses (como o Teste de Bartlett), mas essa matéria não será leccionada.

Uma advertência

Na formulação clássica do modelo ANOVA a dois Factores, com interacção, e a partir da equação-base $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, em vez de impor as condições $\alpha_1 = \beta_1 = (\alpha\beta)_{i1} = (\alpha\beta)_{1j} = 0$ $(\forall i, j)$, admitem-se as restrições:

- $\sum_i \alpha_i = 0$;
- $\sum_j \beta_j = 0$;
- $\sum_i (\alpha\beta)_{ij} = 0\ , \qquad \forall j$;
- $\sum_j (\alpha\beta)_{ij} = 0\ , \qquad \forall i$.

Estas condições alternativas:

- mudam a forma de interpretar os parâmetros;
- mudam os estimadores dos parâmetros;
- não mudam o resultado dos testes F à existência de efeitos.

Delineamentos factoriais com vários factores

Um delineamento factorial (isto é, com observações para todas as combinações de níveis de cada factor) pode ser definido com qualquer número de factores.

Num delineamento factorial a três factores – A, B e C – cada observação da variável resposta indexa-se com quatro índices: Y_{ijkl} indica a observação *l* no nível *i* do Factor A, nível *j* do Factor B e nível *k* do Factor C. A equação de base para Y_{ijkl} prevê a existência de sete tipos de efeitos:

- três efeitos principais de cada factor, α_i, β_j e γ_k.
- três efeitos de interacção dupla associados a cada combinação de níveis de dois Factores diferentes: $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$ e $(\beta\gamma)_{jk}$.
- um efeito de tripla interacção para as células onde se cruzam níveis dos três factores: $(\alpha\beta\gamma)_{ijk}$

O modelo factorial a três factores

A equação de base do modelo é agora:

$$Y_{ijkl} = \mu_{111} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl} \ .$$

A Soma de Quadrados Total é decomposta em oito parcelas: *SQA*, *SQB*, *SQC*, *SQAB*, *SQAC*, *SQBC*, *SQABC* e *SQRE*, de forma análoga ao visto antes.

Os graus de liberdade associados a cada tipo de efeito generalizam conceitos anteriores.

Há sete testes: um para cada tipo de efeitos. As estatísticas desses sete testes são todas do tipo $\frac{QMx}{QMRE}$, onde x designa o tipo de efeitos em questão.

As estatísticas desses testes terão, sob H_0, distribuição F com graus de liberdade dados pelos g.l. do numerador e do denominador, respectivamente.

Delineamentos hierarquizados

Delineamentos que, superficialmente, podem confundir-se com os delineamentos factoriais são delineamentos com dois (ou mais) factores, mas em que os níveis de um dos factores variam consoante os níveis do outro factor.

Exemplo (do Segundo Teste, 2008/9): pretende-se estudar o índice de desempenho (variável resposta), em várias tarefas, de três tractores de diferentes modelos (factor A), cada um dos quais é conduzidos por quatro tractoristas (factor B).

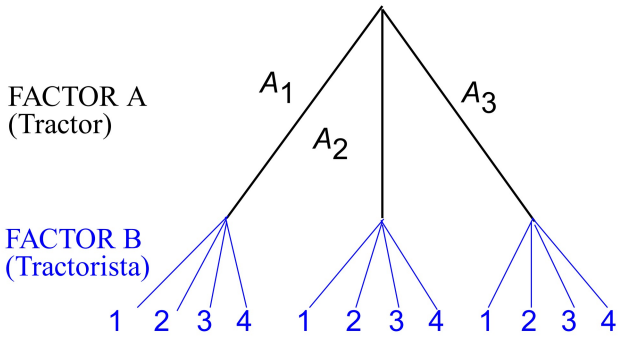
Se os mesmos 4 tractoristas conduzirem os 3 tractores, o delineamento é factorial e aplicam-se os modelos antes considerados.

Mas se para cada modelo de tractor existir um grupo de quatro diferentes tractoristas especializados (ao todo 12 tractoristas), o delineamento não é factorial, mas antes hierarquizado: só é possível identificar os tractoristas (níveis do factor B), após especificar o tractor (nível do factor A).

Delineamentos hierarquizados (cont.)

Existe uma hierarquia dos factores: só identificamos os níveis de um factor (factor subordinado) após ter identificado o nível do outro factor (factor dominante) com que se trabalha.

	Tractor A_1	Tractor A_2	Tractor A_3
Tractorista A_1 1	×	-	-
Tractorista A_1 2	×	-	-
Tractorista A_1 3	×	-	-
Tractorista A_1 4	×	-	-
Tractorista A_2 1	-	×	-
Tractorista A_2 2	-	×	-
Tractorista A_2 3	-	×	-
Tractorista A_2 4	-	×	-
Tractorista A_3 1	-	-	×
Tractorista A_3 2	-	-	×
Tractorista A_3 3	-	-	×
Tractorista A_3 4	-	-	×

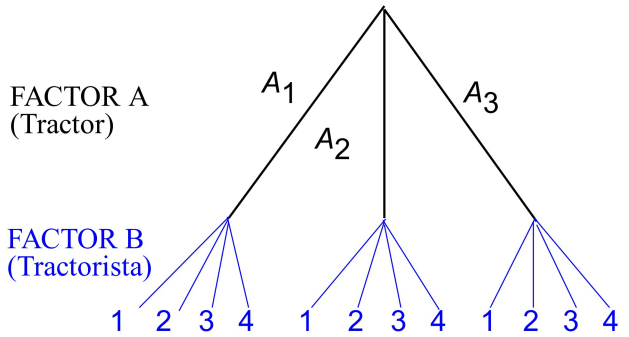


Um tal delineamento diz-se hierarquizado (*nested*, em inglês).

Delineamentos hierarquizados (cont.)

Existe uma hierarquia dos factores: só identificamos os níveis de um factor (factor subordinado) após ter identificado o nível do outro factor (factor dominante) com que se trabalha.

	Tractor A_1	Tractor A_2	Tractor A_3
Tractorista A_1 1	×	-	-
Tractorista A_1 2	×	-	-
Tractorista A_1 3	×	-	-
Tractorista A_1 4	×	-	-
Tractorista A_2 1	-	×	-
Tractorista A_2 2	-	×	-
Tractorista A_2 3	-	×	-
Tractorista A_2 4	-	×	-
Tractorista A_3 1	-	-	×
Tractorista A_3 2	-	-	×
Tractorista A_3 3	-	-	×
Tractorista A_3 4	-	-	×



Um tal delineamento diz-se hierarquizado (*nested*, em inglês).

Um delineamento hierarquizado pode ser visto como um delineamento factorial (muito) incompleto. Deixa de fazer sentido falar em efeitos de interacção entre os níveis de cada Factor.

O modelo a 2 Factores, hierarquizados

Seja b_i o número de níveis do Factor B (folhas terminais do dendrograma), subordinados ao nível i do Factor A (ramo). b_i pode ser diferente para cada nível i do factor dominante.

Cada observação é representada por uma v.a. com três índices, Y_{ijk}:

- i nível do factor dominante ($i = 1, ..., a$);
- j nível do factor subordinado ($j = 1, ..., b_i$);
- k repetição para a célula (i,j), com $k = 1, ..., n_{ij}$.

A equação base do modelo inclui efeitos de nível do Factor A e efeitos de nível do factor B (subordinado):

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk} \ ,$$

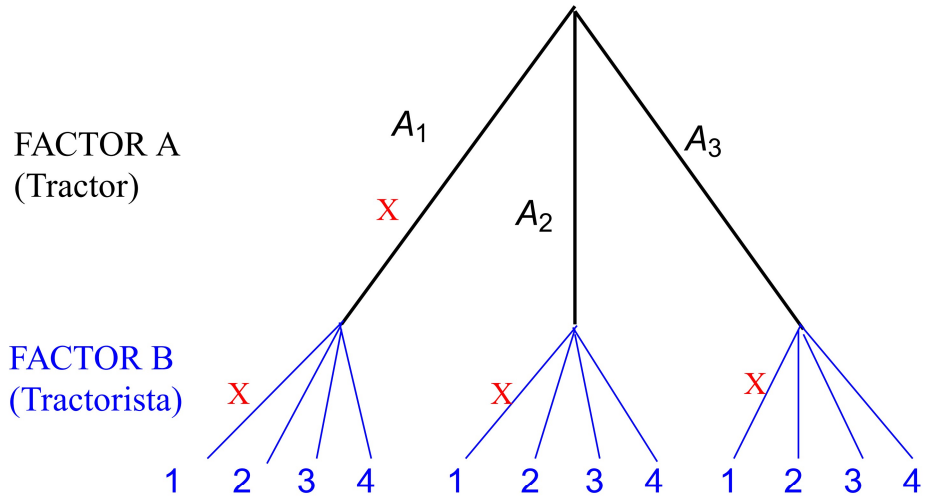
com $\alpha_1 = 0$ e $\beta_{1(i)} = 0$, $\forall i$. Com estas restrições, $\mu = \mu_{11}$.

Não faz sentido falar em efeitos do nível j do Factor B, sem especificar qual o nível do Factor A a que nos referimos. Nem faz sentido falar em efeitos de interacção.

Restrições nos delineamentos hierarquizados

Cada ramo associado ao Factor dominante excepto o primeiro tem efeito α_i.

Cada folha terminal associada ao Factor subordinado excepto a primeira de cada ramo tem efeito $\beta_{j(i)}$.



Os valores esperados de Y_{ijk}

Tem-se:

- Para a primeira célula ($i = j = 1$): $E[Y_{11k}] = \mu = \mu_{11}$.
- Nas restantes células do primeiro nível do Factor A ($i = 1; j > 1$): $\mu_{1j} = E[Y_{1jk}] = \mu_{11} + \beta_{j(1)}$.
- Nos restantes primeiros níveis do factor B ($i > 1; j = 1$): $\mu_{i1} = E[Y_{i1k}] = \mu_{11} + \alpha_i$.
- Nas células genéricas (i, j), com $i > 1$ e $j > 1$, $\mu_{ij} = E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_{j(i)}$.

Os efeitos α_i e $\beta_{j(i)}$ designam-se efeitos dos níveis de cada Factor.

Variáveis indicatrizes e número de parâmetros

Como em modelos anteriores, a cada parâmetro associa-se uma variável indicatriz das observações correspondentes. Assim:

- um parâmetro μ_{11}, associado à coluna de uns, $\vec{\mathbf{1}}_n$.
- $(a-1)$ parâmetros α_i, associados às indicatrizes $\vec{\mathscr{I}}_{A_i}$ de cada nível $i > 1$ do Factor A.
- $\sum_{i=1}^{a} (b_i - 1)$ parâmetros $\beta_{j(i)}$, associados às indicatrizes $\vec{\mathscr{I}}_{B_{j(i)}}$ de cada nível $j > 1$ do Factor B, para $i = 1, ..., a$.

O no. de parâmetros é igual ao no. de situações experimentais:

$$1 + (a-1) + \sum_{i=1}^{a} (b_i - 1) = \cancel{1} + \cancel{a} - \cancel{1} + \sum_{i=1}^{a} b_i - \underbrace{\cancel{\sum_{i=1}^{a} 1}}_{=a} = \sum_{i=1}^{a} b_i$$

Se houver sempre $b = b_i$ níveis do Factor B, em cada nível i do Factor A, haverá ab parâmetros no modelo.

O modelo ANOVA a dois factores, hierarquizados

Juntando os pressupostos necessários à inferência,

Modelo ANOVA a dois factores, hierarquizados (Modelo $M_{A/B}$)

Seja *A* o Factor dominante e *B* o Factor subordinado.
Existem n observações, Y_{ijk}, n_{ij} das quais associadas à célula (i,j) $(i = 1,...,a\ ;\ j = 1,...,b_i)$. Tem-se:

1. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}\ , \quad \forall i{=}1,...,a\ ;\ j{=}1,...,b_i\ ;\ k{=}1,...,n_{ij}$
   $(\alpha_1 = 0\ ;\ \beta_{1(i)} = 0\ ,\ \forall i)$.
2. $\varepsilon_{ijk} \frown \mathcal{N}(0, \sigma^2)\ , \qquad \forall i,j,k$
3. $\{\varepsilon_{ijk}\}_{i,j,k}$ v.a.s independentes.

Os dois testes ANOVA

Neste delineamento, pretende-se testar a existência de cada um dos dois tipos de efeitos previstos no modelo:

- $H_0 : \alpha_i = 0\,, \quad \forall i = 2, ..., a$; e
- $H_0 : \beta_{j(i)} = 0\,, \quad \forall i = 1, ..., a$ e $j = 2, ..., b_i$.

As estatísticas de teste para cada um destes testes obtêm-se a partir da decomposição da Soma de Quadrados Total em três parcelas, correspondentes aos dois tipos de efeito e à variabilidade residual.

As Somas de Quadrados associadas a cada tipo de efeito definem-se de forma análoga à usada em delineamentos anteriores.

A decomposição de SQT

Para efectuar a decomposição da Soma de Quadrados Total, consideremos os modelos

$$\begin{aligned} (\text{Modelo } M_{A/B}) \qquad & Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk} \ , \\ (\text{Modelo } M_A) \qquad & Y_{ijk} = \mu_{11} + \alpha_i + \varepsilon_{ijk} \ , \end{aligned}$$

Designa-se Soma de Quadrados associada aos efeitos de B a

$$SQB(A) = SQRE_A - SQRE_{A/B}$$

e Soma de Quadrados associada aos efeitos de A a

$$SQA = SQF_A = SQT - SQRE_A$$

Juntamente com $SQRE_{A/B}$, tem-se:

$$SQT = SQA + SQB(A) + SQRE_{A/B}$$

Algumas fórmulas

Como $SQA = SQF_A$ (Modelo 1 Factor):

$$SQA = \sum_{i=1}^{a}\sum_{j=i}^{b_i}\sum_{k=1}^{n_{ij}}(\underbrace{\hat{Y}_{ijk}}_{=\overline{Y}_{i..}} - \overline{Y}_{...})^2 = \sum_{i=1}^{a}\sum_{j=i}^{b_i} n_{ij}\,(\overline{Y}_{i..} - \overline{Y}_{...})^2 \,.$$

Num delineamento equilibrado, tem-se: $SQA = n_c \sum_{i=1}^{a} b_i\,(\overline{Y}_{i..} - \overline{Y}_{...})^2$

No modelo a 2 factores hierarquizado também se tem:

$$\hat{Y}_{ijk} = \overline{Y}_{ij.}$$

Logo, a Soma de Quadrados Residual também é soma ponderada das variâncias de célula $S_{ij}^2 = \frac{1}{n_{ij}-1}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \overline{Y}_{ij.})^2$:

$$SQRE = \sum_{i=1}^{a}\sum_{j=i}^{b_i}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \underbrace{\hat{Y}_{ijk}}_{=\overline{Y}_{ij.}})^2 = \sum_{i=1}^{a}\sum_{j=i}^{b_i}(n_{ij}-1)\,S_{ij}^2 \,.$$

Graus de liberdade

Os graus de liberdade associados a cada tipo de efeito são dados por:

- $g.l.(SQA) = a - 1$, o número de parâmetros associados aos efeitos de nível de A.
- $g.l.[SQB(A)] = \sum_{i=1}^{a} (b_i - 1)$, o número de parâmetros associados aos efeitos de nível de B.
- $g.l.(SQRE) = n - \sum_{i=1}^{a} b_i$, o número de observações menos o número total de parâmetros do modelo.

Quadro-resumo da ANOVA a 2 Factores hierarquizados

Fonte	g.l.	SQ	QM	f_{calc}
Factor A	$a-1$	SQA	$QMA = \frac{SQA}{a-1}$	$\frac{QMA}{QMRE}$
Factor B(A)	$\sum_{i=1}^{a} (b_i - 1)$	$SQB(A)$	$QMB(A) = \frac{SQB(A)}{\sum_{i=1}^{a} (b_i - 1)}$	$\frac{QMB(A)}{QMRE}$
Resíduos	$n - \sum_{i=1}^{a} b_i$	$SQRE$	$QMRE = \frac{SQRE}{n - \sum_{i=1}^{a} b_i}$	
Total	$n-1$	$SQT = (n-1)\,S_y^2$	–	–

O Teste F aos efeitos do factor A (dominante)

Sendo válido o Modelo de ANOVA a 2 factores hierarquizados, tem-se:

Teste F aos efeitos do factor A (dominante)

Hipóteses: $H_0 : \alpha_i = 0 \quad \forall i=2,...,a$ vs. $H_1 : \exists i=2,..,a \text{ t.q. } \alpha_i \neq 0$.

[FACTOR A NÃO AFECTA] vs. [FACTOR A AFECTA Y]

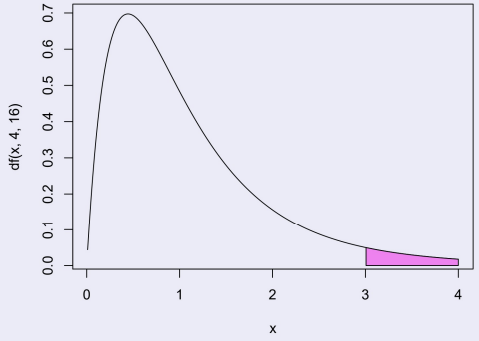
Estatística do Teste: $F = \frac{QMA}{QMRE} \frown F_{(a-1\,,n-\sum_i b_i)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha(\,a-1\,,n-\sum_i b_i\,)}$



O Teste F aos efeitos do factor B (subordinado)

Sendo válido o Modelo de ANOVA a dois factores hierarquizado,

Teste F aos efeitos do factor B (subordinado)

Hipóteses: $H_0 : \beta_{j(i)} = 0 \quad \forall j=2,\ldots,b_i \ , \ i=1,\ldots,a$ vs. $H_1 : \exists i,j \text{ t.q. } \beta_{j(i)} \neq 0$.

[FACTOR B NÃO AFECTA] vs. [FACTOR B AFECTA Y]

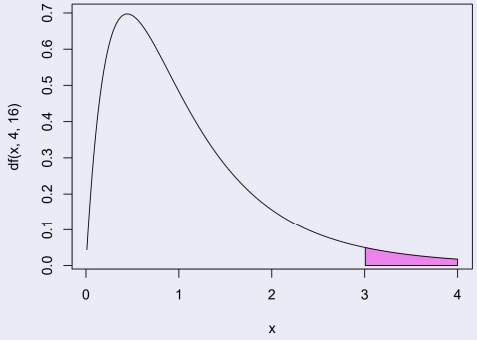
Estatística do Teste: $F = \frac{QMB(A)}{QMRE} \frown F_{(\sum_i (b_i-1),\, n-\sum_i b_i)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha(\sum_i (b_i-1),\, n-\sum_i b_i)}$



ANOVA a dois Factores hierarquizados no R

Para efectuar uma ANOVA a dois Factores hierarquizados no R, organizam-se os dados como nos anteriores modelos com dois factores, ou seja, numa `data.frame` com três colunas:

1. uma para a variável resposta;
2. outra para o factor A;
3. outra para o factor B.

A fórmula utilizada no R para indicar uma ANOVA a dois Factores hierarquizados é semelhante às anteriores, mas com o nome dos dois factores separado pelo símbolo /. Se o factor *fA* é dominante:

```
y  ~  fA / fB
```

Um exemplo

Exemplo de delineamento hierarquizado

No exemplo de tractores/tractoristas, o delineamento era equilibrado, com $n_c = 5$ observações em cada célula (situação experimental).

A tabela-resumo produzida pelo comando `aov` é a seguinte:

```
> summary(aov(indice ~ tractor/tractorista, data=tractores))
                    Df Sum Sq Mean Sq F value   Pr(>F)
tractor              2   1696   847.8   35.92 2.90e-10 ***
tractor:tractorista  9   2272   252.5   10.70 6.99e-09 ***
Residuals           48   1133    23.6
```

Neste caso, há efeitos significativos dos diferentes tipos de tractores sobre a variável resposta, e também efeitos significativos dos tractoristas que conduzem os tractores.

Comparações múltiplas de médias

Caso se conclua pela existência de efeitos do factor subordinado, é natural querer comparar médias da variável resposta nas $\sum_{i=1}^{a} b_i$ diferentes situações experimentais.

Comparações múltiplas de Tukey podem ser efectuadas, caso o delineamento seja equilibrado, isto é, se houver o mesmo número de observações em cada situação experimental.

Neste caso, os parâmetros da distribuição de Tukey serão

- o número de situações experimentais, $k = \sum_{i=1}^{a} b_i$; e
- os graus de liberdade associados ao $QMRE$, $\nu = n - \sum_{i=1}^{a} b_i$.

Tukey – Um exemplo

Tukey com os dados dos tractoristas

Há $b_1 + b_2 + b_3 = 12$ situações experimentais, logo $\binom{12}{2} = 66$ comparações de pares de médias dessas situações experimentais. O termo de comparação de Tukey para diferenças de médias de célula é:

$$q_{0.05\,(12,48)} \cdot \sqrt{\frac{QMRE}{n_c}} = 4.856029 \times \frac{4.85793}{\sqrt{5}} = 10.55$$

As médias de célula são:

```
> model.tables(tractores.aov, type="means")
[...]
       tractorista
tractor 1    2    3    4
      1 61.8 67.8 62.6 52.6
      2 75.8 75.2 55.8 77.0
      3 76.8 69.6 74.4 73.4
```

O maior índice médio de desempenho é $\overline{y}_{24.} = 77.0$. A azul estão as médias que não diferem significativamente da maior média. A preto ficam as que diferem.

Validação do modelo

A análise de resíduos para validar os pressupostos do modelo, é análoga à de modelos anteriores.

Gráficos de resíduos no exemplo dos tractores

```
> plot(tractores.aov, which=c(1,2))
```

